

THIS WEEK

EDITORIALS

FEYNMAN Fifty years of those famous physics lectures and textbook **p.8**

WORLD VIEW A target of sexual harassment speaks out **p.9**



STEALTH Snout takes sea horse smoothly through water **p.11**

Call the cops

The long arm of the law has reached into an investigation of alleged scientific misconduct in Italy, and should perhaps stretch still farther.

Science likes to shelter its crooks with euphemisms. The prefix ‘research’ softens fraud, and to deliberately obtain public money through deception gets labelled misconduct, among other things. This reflects the fact that the crime is viewed as being against professional standards rather than against the laws of wider society. The international game of science turned professional long ago, but the rules of play and their enforcement still harbour a decidedly amateur spirit.

Germany, for instance, lost its science-integrity innocence some 15 years ago when two high-flying biomedical researchers, Friedhelm Herrmann and Marion Brach, were found to have manipulated or fabricated data in 94 published papers. The affair — which became a world reference case for scientific fraud — shattered the complacency of German academia, which had never imagined that such bad conduct could occur in its world, and had no mechanism for dealing with it. In response, it established an admirable system of self-regulation, arming universities with clear guidelines on good scientific practice, procedures for investigating allegations of misconduct and an ombudsman system to which whistle-blowers can turn. It works well, if rather slowly.

Most scientifically ambitious countries now have similar systems in place, frequently rustled up in response to their own local scandals. The procedural details vary, but all the systems aim to protect the whistle-blower and the accused during investigations and, crucially, carry out those investigations within the academic community. Self-regulation is written into the academic scriptures: what outsider could possibly understand the nuances of the scientific system, with its fundamental requirement for freedom of inquiry? But a case of alleged scientific misconduct currently under investigation in Italy begs a question, perhaps a heretical one: could the police sometimes do a better job at getting to the truth than the academic community?

Italy has no scientific-misconduct procedures in place, so last year, a frustrated whistle-blower presented his allegations of inappropriate duplication against a cancer researcher to the police (see page 18). The police seem to be carrying out a serious, detailed and thoroughly professional investigation. (No charges have been brought.)

Scientists, of course, should not have to live with the threat of being marched out of their lab in handcuffs whenever casual allegations are made. But, as the case in Italy shows, that is not how it has to work. The case is months from reaching a conclusion, but it is already clear that the procedures used by the police are no less protective of the parties involved than an academic inquiry would have been. Details were eventually leaked to the press, but only 20 months after the investigation began, and even then the whistle-blower was not exposed. The police are also at least as professional as academics, and, with their powers to confiscate computers and laboratory notebooks, can help the investigation to move forward more effectively. Academic investigators are often frustrated by accused scientists refusing to provide such evidence.

A natural response to a police investigation is that outsiders could never understand the academic system well enough to sit in judgement.

Really? Police forces worldwide routinely deal with financial and computer crimes, the details of which can seem equally impenetrable. Understanding what a western blot is and why it shouldn't be tampered with are obvious challenges for a non-scientist — as is understanding the mysteries of the academic world and the role of peer-reviewed publications within it. But the police know a thing or two about conducting an investigation. And any external inquiry has a distinct advantage: it

“Academic investigators could learn from police methods.”

cannot be hindered by the intrinsic threat of conflict of interest that comes when any community sits in judgement on its own members.

Admittedly, the learning curve for the Italian public prosecutors in this case has been extraordinarily steep. But the police can easily work out who should be called in to instruct them. In fact, they must be careful not to get things wrong, or they will be humiliated should a case reach court.

Science, of course, might not seem a top priority for law enforcers. And it is true that not all police departments would have the mindset to take fraud in arcane areas of science very seriously. Although much public money can be lost, fraudulent science must compete with more familiar crimes, such as tax evasion (and it pales in comparison with violent crimes).

Still, the Italian example does deserve broader discussion. At the very least, academic investigators could learn from police methods for dealing with allegations of serious misconduct — that word again. And researchers might be less tempted to be cavalier with the truth — and with our money — if they knew who else could knock on their door. ■

The FDA and me

Medical testing firms find it is in their best interests to cooperate with regulators.

Late last month, US regulators dropped a bombshell on the genetic-testing start-up 23andMe in an exasperated cease-and-desist letter that prompted a fast and contrite response from the company — and a flurry of criticism of both parties among scientists and self-styled Health 2.0 activists who advocate the use of Internet tools in medicine.

Since 2007, 23andMe, which is based in Mountain View, California, has been testing customers' DNA for a range of traits, from the frivolous, such as earwax type, to the more significant, such as disease risk and genetic ancestry. The company has walked a fine line between promising that this activity will revolutionize medicine and averring

that it is not actually medical at all, in an attempt to simultaneously lure in customers and avoid the need to conform to medical regulations.

The US Food and Drug Administration (FDA) has now called 23andMe's bluff, complaining that the company has "not completed" some studies that would prove the soundness of its methods and "not even started" others; that 23andMe has shunned communication with the FDA since May; and that the company has launched a large advertising campaign without getting marketing approval. The agency demanded that 23andMe stop marketing its testing kit until it received proper authorization.

The episode has been interpreted as everything from a massive regulatory overreach that threatens to quash innovation, to a long-needed dose of supervision for a dangerously out-of-control industry.

But the big question is not whether regulators will stop people from understanding their own DNA — they cannot. The question is whether such understanding has reached the point at which companies can exploit it, and if so, how to protect their customers. Part of answering that question is determining whether a company's claim is true. This is what the FDA is trying to do, and until earlier this year, it seemed that 23andMe was happy to aid that mission — FDA approval, after all, would dispel worrying chatter about whether regulators would ultimately shut the company down. Mainstream biotechnology companies learned a long time ago that it pays to play nice with regulators.

It is unclear whether 23andMe's six-month lapse in communication with the FDA stems from inexperience with regulatory procedures, or from a hope that it could quickly grow its customer base large enough to monetize in other ways. The problem with the latter strategy is that direct-to-consumer medical genetic testing is not yet a viable business model.

The company's chief executive, Anne Wojcicki, told a conference at Stanford University in California in May that 23andMe hoped to amass 1 million customers by the end of this year, but the company still has only half that number. And other firms in the market have not succeeded: last year, Navigenics of Foster City, California, was acquired by biotech firm Life Technologies and stopped offering consumer testing, and deCODEme of Reykjavik shut down.

Consumer demand is low in part because genetic tests on healthy people still cannot be relied on to produce consistent predictions about medical risks. Customers of 23andMe have detailed how the service variously provides lifesaving information and misleading results. This is simply the state of the science today. Silicon Valley 'health disrupters' who plan to revolutionize health care, such as Wojcicki and her estranged husband, Google co-founder Sergey Brin, like to think that

"Direct-to-consumer medical genetic testing is not yet a viable business model."

they can apply their successful data-mining strategies to medicine, but it turns out that biology is more complicated than they perhaps first assumed.

No one should be fooled into thinking that direct-to-consumer genetic testing is doomed to fail. The science is moving so much faster than medical education that motivated and self-taught laypersons can learn and understand just as much about their genetic medical risks as can their doctors. Indeed, there are already public crowd-sourced tools that customers can use to interpret their genetic data for free. So even if regulators or doctors want to, they will not be able to stand between ordinary people and their DNA for very long.

In the meantime, it seems short-sighted for companies to rebuff regulators. If it is too onerous to prove the accuracy of the information they offer, they should not be selling this information in the first place. And if they turn up their noses at regulators, they may run afoul of an even more powerful force: the US system of civil litigation. Consumers are already joining class-action lawsuits alleging that 23andMe is selling misleading information. Such suits are much more effective than anything the government can do to get companies to change their practices.

To its credit, 23andMe seems to have learned this: on 26 November, Wojcicki acknowledged in a blog post both that the "FDA needs to be convinced of the quality of our data" and that "we are behind schedule with our responses" to the agency. The company has also stopped marketing.

It seems, then, that 23andMe's experience with the FDA is less about the growing pains of a new industry than about affirming a principle — the need for truth in advertising — that is as old as business itself. ■

Lecture notes

A physics course that hooked a generation reminds us that teachers need support.

It's a 50-year-old physics textbook that runs to 1,500 pages and whose contents were declared a failure by its famous author. It is also, according to various online reviews "spellbinding" and "an extraordinary book written by an extraordinary man". One goes as far as to say: "Here's the deal. If ya wanna do this whole physics thing vanilla-style, go buy and read a nice physics textbook. If you want to taste physics — really take it in, like a delicious chocolate mousse or a symphony orchestra or Shakespeare done by British folk, this is where you have to be."

Perhaps the most extraordinary thing about *The Feynman Lectures on Physics*, the book in question, is that it was nearly strangled at birth. Robert Leighton, chair of a committee tasked with spicing up the physics teaching at the California Institute of Technology in Pasadena in the early 1960s, did not think that Richard Feynman was the right man for the job. "That's not a good idea," was his original response. "Feynman has never taught an undergraduate course. He wouldn't know how to speak to freshmen, or what they could learn." (At around the same time, incidentally, an official at Decca Records decided that "The Beatles have no future in show business".)

Leighton was won round, but the transition from a limited series of lectures — delivered only once by Feynman, between 1961 and 1963

— to a textbook that still inspires devotion five decades on was equally hesitant. As Matthew Sands, who helped to organize the lectures and is a co-author on the book, recalled in 2005, the first draft received from the publishers was a "disaster" (M. Sands *Phys. Today* **58**, 49–55; April 2005). A well-meaning editor had rewritten Feynman's informal style into more traditional textbook-speak; notably, the physicist's conversational 'you' had been inelegantly changed to 'one'. (Sands also recalled Feynman's first reaction to the idea that he would share authorship credit with Sands and Leighton: "Why should your names be there? You were only doing the work of a stenographer!")

As Rob Phillips explores in an In Retrospect article on page 30 of this issue, *The Feynman Lectures* has endured because it was ahead of its time, and because "his introduction to elementary physics seems to have higher aspirations — the love of nature and a grasp of it through experimentation and reasoning". In Feynman's hands, physics turned from a description of the world to a way of thinking about it, and a generation was hooked.

The popularity of the lectures and the enduring appeal of the books that grew from them are often attributed to the individual and spontaneous genius of Feynman. But they were painstakingly prepared and practised, and had generous financial backing. (The lectures were part of broader changes to the teaching at Caltech's physics department funded with some US\$1 million from the Ford Foundation.)

This is a lesson that university officials would do well to remember as funding is cut and pressure placed on faculty members to cram more into their timetables. Those who can, teach, but they need support. ■

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunq

JESSICA HORWITZ



How sexual harassment changed the way I work

As a flurry of interest in workplace discrimination subsides, efforts to raise awareness and eliminate abuses continue, says **Kathleen Raven**.

In 2012, my male editor suggested that he would like to have sex with me. I politely declined. Many apologies later, he explained that he really just wanted to be friends. A friendship seemed feasible yet frightening. We worked for the same company, and a power differential existed between us. I told him as much, not that it helped. For a year and a half, as often as I defined my boundaries, he trampled over them. One by one, I shut down our avenues of communication. “Who is this guy who ‘likes’ everything you write on Facebook?” my parents asked. I quit Facebook. But to remove myself from the situation would have meant leaving my chosen profession, science writing.

Women who choose careers in male-dominated domains such as politics, engineering, publishing, business, mathematics, computer science and science writing know that sexual harassment is par for the course. Unlike most of them, I did not keep quiet. As a simple Google search reveals, the editor who harassed me was Bora Zivkovic, who resigned from his position as blogs editor of *Scientific American* after I and other women complained publicly about his behaviour.

After I spoke out, other women privately shared with me their stories of harassment by men. Their relief felt palpable. On Twitter, men and women responded with overwhelming support. Yet at the meeting of the US National Association of Science Writers in Gainesville, Florida, last month, I felt anxious when I spoke at a panel dedicated to raising awareness of how women still struggle in the industry: how far behind they remain in terms of winning awards or getting blogging gigs with leading publishing brands. I did not enjoy being the public face of a flurry of interest in the problem of sexual harassment.

Even the phrase ‘sexual harassment’ is dangerous. Men may run away in fear and assume that any interaction with women can be wrongly interpreted. Women may not know that they are protected against certain actions, and may be nervous about speaking out because of the implications for themselves or for other women.

A few mostly anonymous commenters on Twitter and in the blogosphere have criticized my method of coming forward. I never doubt I did the right thing, but I do not want to set a precedent. History has never looked kindly on witch hunts. Today’s self-publishing environment means that defamation can be only a few keystrokes away. My public comments — in a blog post I wrote — could have been avoided. I realize now that if I challenge an offender either directly or through confidential official channels, then both women and men will listen and look out for me. I did not know that before.

I have endured various types of sexual harassment throughout my career, from the relatively

harmless to the illegal. When I was 16, the editor at my first newspaper job told me he loved me. His habit of stroking my calf muscles communicated that he meant this romantically — not as a compliment on my reporting. A male executive of a different publishing company habitually stopped by my desk, to lightly massage my shoulders and tell me: “What a great smile you have!”

Zivkovic was certainly not the first. But I can say with confidence that he was the last. The experience changed the way I interact with men in professional settings. I am ready to tell them immediately if they step over a line. And if that does not work, then I will simply walk away. The situation has provided me a sort of shield, too, in the way that men may interact with me. If I am labelled as the ‘woman who writes about sexual harassment’ and kept at arm’s length, then so be it.

By now, many in the science community have moved on from discussions of this problem. But the work of raising awareness of and eliminating sexual harassment has not stopped. It continues very vividly in a core, but growing, group of female science writers whom I know and admire. And I continue to ask my female colleagues about their workplace environments. For example, I helped one of my colleagues to send a letter to her supervisor, who was making quite subtle sexual comments. After the letter, the comments stopped. Women and men can seek such small victories every day.

Throughout the conversation, some have claimed that modern women are being overly sensitive. Do we read stories about Marie Curie, Rosalind Franklin or, more recently, Rita Levi-Montalcini or Mae Jemison complaining about sexual harassment? We do not. But for these women, legal protection did not exist or had only just begun.

The law in the United States now is simple enough: “Harassment can include ‘sexual harassment’ or unwelcome sexual advances, requests for sexual favors, and other verbal or physical harassment of a sexual nature,” states the US Equal Employment Opportunity Commission on its website. The real work begins when a woman is confronted with a situation and must define it. (I focus on women here because, out of the nearly 11,400 sexual-harassment complaints filed to the commission in 2011, only about 2,000 came from men.) That definition can be boiled down to a simple rule that men must follow when they interact with female colleagues. Ask the question: “Would I say this to a man?” ■

Kathleen Raven is a freelance science journalist in Atlanta, Georgia. Find her on Twitter @sci2mrow. e-mail: kathraven@gmail.com

WOMEN MAY NOT
KNOW THAT
**THEY ARE
PROTECTED,**
AND MAY BE NERVOUS
ABOUT
**SPEAKING
OUT.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/qvzblc

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

GEOLOGY

Clues to extinction in lava gases

Acid rain and ozone depletion probably contributed to the greatest extinction Earth has ever seen.

Massive volcanic eruptions occurred in Siberia at around the same time as the extinction event that ended the Permian era some 252 million years ago, but it is not clear how the two are linked. Benjamin Black of the Massachusetts Institute of Technology in Cambridge and his colleagues analysed the amounts of gases trapped in the Siberian lava. They put the data into a global climate model describing the ancient atmosphere.

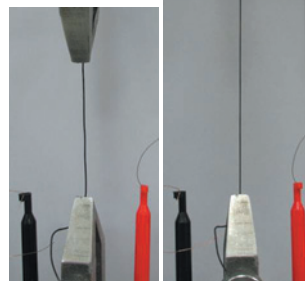
The levels of some gases emitted by the magma, such as carbon dioxide and sulphur dioxide, would have turned rainfall to acid. Others, such as methyl chloride, may have chewed away at the planet's protective ozone layer.

Geology <http://doi.org/p7s> (2013)

ELECTRONICS

Power storage in stretchy fibres

Wearable electronics require stretchy components, but such parts are often flat sheets, which limits their ability to be incorporated into fabrics. So Huisheng Peng and his colleagues at Fudan University in Shanghai, China, have developed supercapacitors



— devices that store electric charge — in the form of highly stretchable fibres, which can be woven into textiles.

The researchers made the supercapacitors by coating rubber fibres with alternating layers of an electrolyte and sheets of carbon nanotubes that act as electrodes. Even when the fibres are stretched more than 100 times to 175% (pictured) of their original length, they still provide a power output that is equivalent to other carbon-based, fibre-shaped supercapacitors that are not stretchy.

Angew. Chem. Int. Edn <http://doi.org/f2nqcn> (2013)

ECOLOGY

Sharks never forget home

Female lemon sharks return to their birth waters to deliver offspring — the first direct observation of such behaviour in any shark species.

Kevin Feldheim at the Field Museum in Chicago, Illinois, and his colleagues collected and analysed DNA from lemon sharks (*Negaprion brevirostris*; pictured) in the waters surrounding the Bimini Islands in the Bahamas every year from 1995 to 2012. At least six

female sharks born there between 1993 and 1998 later returned to give birth. The females were also faithful to particular nursery areas between the islands.

Conservation efforts should limit fishing in such areas when females return to give birth, or should look to establish marine reserves that encompass the nurseries, the researchers suggest. **Mol. Ecol.** <http://doi.org/p78> (2013)



CANCER

How cholesterol drives tumours

A cholesterol breakdown product speeds up the rate at which tumours grow and spread in mouse models of breast cancer.

High cholesterol is a known risk factor for breast cancer, so Donald McDonnell at Duke University in Durham, North Carolina, and his colleagues increased blood levels of a cholesterol metabolite, 27HC, in mice with breast tumours. Tumours in treated mice

grew faster and spread to the lungs more often than in untreated animals. In human cancer cells, higher expression of an enzyme that converts cholesterol to 27HC correlated with more-aggressive tumours. In one mouse model, a high-fat diet boosted 27HC levels in the blood and increased tumour growth, but this slowed down when the animals were given a cholesterol-lowering statin.

The breakdown product drives tumour growth by binding to the receptor for oestrogen. **Science** 342, 1094–1098 (2013)

NEUROSCIENCE

Hormone boosts attractiveness

The hormone oxytocin may contribute to the romantic bonds that keep men faithful.

René Hurlemann at the University of Bonn, Germany, and his colleagues used functional magnetic resonance imaging to study the brains of two groups of 20 men. During the scans, one group looked at pictures of their female partners and unfamiliar women assessed by the researchers as being equally attractive, and the other group looked at photos of their partners and familiar women who were not relatives.

All 40 men rated their partners as more attractive than either unfamiliar or familiar women. But men who received nasal puffs of oxytocin before scanning gave higher ratings for their partners than did those who received a placebo. This boost occurred only with partners, and not with familiar women.

Men given oxytocin also had increased signalling in the reward centres of the brain when shown their partner's face, but not when shown a picture of a familiar woman. *Proc. Natl Acad. Sci. USA* <http://doi.org/p74> (2013)

CHEMISTRY

Catalysts on the cheap

Chemists are making rapid progress in replacing catalysts that use precious metals, such as platinum and iridium, with catalytic molecules based on more abundant metals. Now, three groups have reported improved methods for adding hydrogen to particular parts of molecules — 'hydrogenation' reactions that are involved in making drugs, polymers and other industrial chemicals.

A team led by Paul Chirik at Princeton University in New Jersey studied catalysts that are based on simple cobalt salts wrapped by other widely

available molecules. These are adept at hydrogenating a variety of carbon-carbon double bonds.

Robert Morris's group at the University of Toronto in Canada created iron-based catalysts that can hydrogenate carbon-oxygen or carbon-nitrogen double bonds. Such catalysts are more active than commercial platinum compounds and are just as selective in producing the desired version of a compound.

And Matthias Beller at the University of Rostock in Germany and his colleagues found that catalysts using solid iron-oxide nanoparticles do well at hydrogenating another chemical structure, the aryl nitro group — useful in agrochemicals and dyes, for instance.

Science 342, 1073–1076; 1076–1080; 1080–1083 (2013)

MOLECULAR BIOLOGY

RNAs leave yeast poised for action

One way in which long, non-coding RNAs control gene expression in yeast is to accelerate the activation of protein-coding genes.

Elizabeth Tran and her colleagues at Purdue University in West Lafayette, Indiana, studied various strains of *Saccharomyces cerevisiae*, in which the *GAL* genes are repressed or activated by different sugars in the environment. The team found that when these genes are released from a repressed state, long non-coding RNA molecules (lncRNAs) speed up *GAL* gene expression by quickly recruiting a key enzyme needed to make proteins. The lncRNAs also hinder the binding of molecules that repress *GAL* genes.

The team suggests that these particular lncRNAs leave yeast poised and ready to quickly switch carbon sources in response to environmental changes.

PLoS Biol. 11, e1001715 (2013)

COMMUNITY CHOICE

The most viewed papers in science

Virology

Chinese origin for US pig virus

HIGHLY READ
on www.asm.org
in October

A mysterious pig virus outbreak that erupted in the United States in May 2013 was imported from China.

A team led by Xiang-Jin Meng and Yao-Wei Huang of Virginia Polytechnic Institute and State University in Blacksburg sequenced the genome of three US strains of porcine epidemic diarrhoea virus (PEDV), which has killed piglets at an alarming rate on hundreds of farms across the country.

The team found that these strains were similar to one found in Anhui province in China in 2010. Furthermore, one section of the PEDV genome is similar to that of a bat virus, suggesting that the pathogen can be transmitted between species.

mBio 4, e00737-13 (2013)

MATERIALS SCIENCE

Sticky surface switches on and off

A magnetic material mimics the gravity-defying stickiness of a gecko's footpads and, notably, the creature's ability to turn this property on and off.

Researchers led by Aránzazu del Campo at the Max Planck Institute for Polymer Research in Mainz, Germany, created an adhesive surface made of arrays of T-shaped micropillars coated with neodymium magnet particles. Applying a magnetic field to this surface bent the pillars and quickly turned off its stickiness. The reversible system works in wet and dry conditions, and it can be easily prepared and scaled up, the authors say.

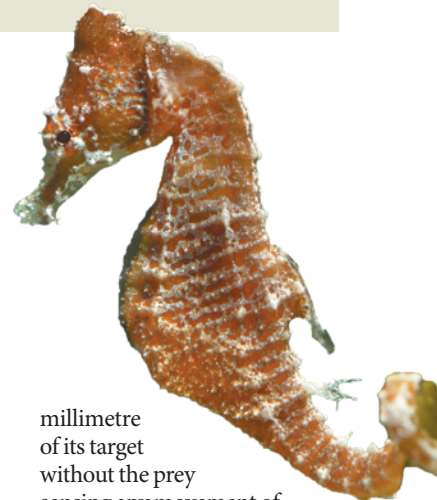
Adv. Mater. <http://doi.org/p8m> (2013)

ZOOLOGY

Stealthy sea horse uses its head

The long snout of the dwarf sea horse (*Hippocampus zosterae*; pictured) allows it to sneak up on its prey.

The creature sucks its victim into its mouth by quickly snapping its head upwards, but to do so it must get within one



millimetre of its target without the prey sensing any movement of the surrounding water.

A team led by Brad Gemmell at the University of Texas at Austin used three-dimensional digital holography to track the flow of water around a sea horse as it swam towards a small crustacean. They found that the unique shape of the sea horse's head, along with its orientation, creates a zone in which water is undisturbed as the sea horse moves.

The results could be relevant for the design of microfluidic devices that need to move water with minimal disturbances.

Nature Commun. 4, 2840 (2013)

NATURE.COM

For the latest research published by Nature visit:

www.nature.com/latestresearch

SEVEN DAYS

The news in brief

POLICY

Israeli settlement

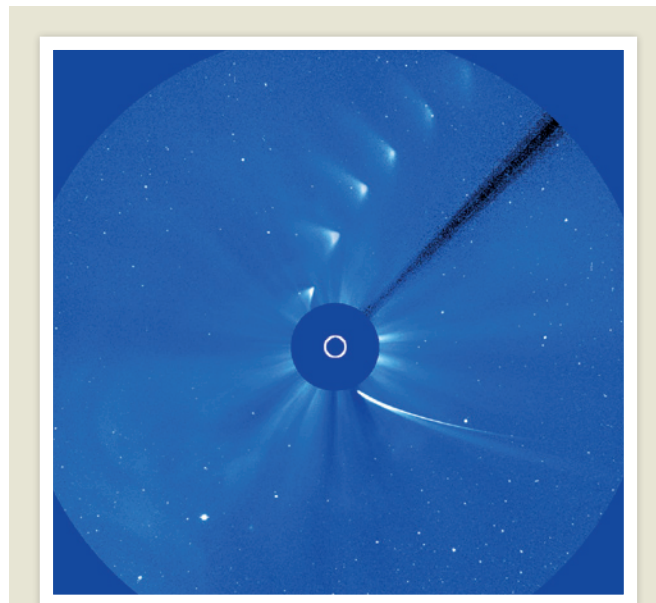
The Israeli government and the European Union (EU) have resolved a political stand-off that had threatened to stop Israeli scientists from participating in Horizon 2020, the EU's €80-billion (US\$109-billion) research-funding programme for 2014–20. EU guidelines banned money from being spent on research in territories outside Israel's pre-1967 borders, a measure that Israel objected to (see *Nature* **501**, 293–294; 2013). On 26 November the country signed up to the programme, but will note in an appendix that it has legal and political objections to the guidelines — a compromise that satisfied both sides.

Drug U-turn

US regulators have removed the safety restrictions that they put on the diabetes drug Avandia (rosiglitazone) in 2010, after it was linked to an increased risk of heart attack. The Food and Drug Administration reversed its stance on 25 November; it said that a reanalysis of a pivotal clinical trial suggested that people taking the drug were not more likely than others to die from heart complications. Avandia, made by GlaxoSmithKline, remains banned in Europe. See go.nature.com/kwnzmj for more.

HIV up in Europe

The number of new HIV infections reported in Europe increased by 8% between 2011 and 2012, the European Centre for Disease Prevention and Control and the World Health Organization said on 27 November. About 78% of the roughly 131,000 new cases were reported in Eastern Europe and central Asia, the groups found, with AIDS



Death of a comet

Comet ISON apparently broke apart as it passed close to the Sun on 28 November, leaving behind what is probably only a cloud of dust. The disintegration disappointed skywatchers across the Northern Hemisphere, who had hoped for a naked-eye view of a bright comet as ISON emerged on the other side of the Sun. Images from the Solar and Heliospheric Observatory (see time-lapse picture) and other Sun-watching satellites showed the comet's brightness and dust production peaking just before it swung past the Sun within a distance of 1.2 million kilometres. ISON was unusual because it was both a first-time visitor to the inner Solar System and a 'sungrazer' comet. See go.nature.com/2fgjtd for more.

diagnoses in Eastern Europe growing by 113% in 2006–12. The organizations say that poor access to preventive measures and to antiretroviral therapy has contributed to the spike.

Science education

China, Singapore, Japan and Finland topped rankings of educational excellence in science among school students, according to the latest Programme for International Student Assessment (PISA) results released on 3 December. The study, which is performed every three years by the Organisation for

Economic Co-operation and Development in Paris, surveys attainment in mathematics, reading and science in 15-year-old students. Boys and girls performed similarly in science.

EVENTS

Non-human rights

On 2 December, a group called the Nonhuman Rights Project filed the first of three US lawsuits intended to grant legal 'personhood' rights to chimpanzees. The Florida-based group is suing on behalf of four chimps in New York state — one suit

involves two chimps at Stony Brook University and the other two are for animals under private ownership. The group charges that the animals are being denied bodily freedom. In May, India's environment ministry banned the use of captive dolphins in entertainment, adding that they should be seen as "non-human persons".

On track to Mars

India's mission to Mars has left Earth's orbit. On 1 December, the Mars Orbiter Mission spacecraft (informally called Mangalyaan) fired its engines and began its 300-day journey to the red planet. Launched on 5 November, the 4.5-billion-rupee (US\$72-million) mission is the country's first interplanetary probe.

RESEARCH

X-ray vision

An X-ray telescope will be the European Space Agency's next large mission, the agency confirmed on 28 November. Scheduled for launch in 2028, the €1-billion (US\$1.4-billion) craft will study how gas evolves into galaxies and how black holes grow and influence their surroundings (see details at *Nature* **503**, 13–14; 2013). The next mission after that, a €1-billion gravitational-wave space observatory, is scheduled for 2034, the agency said. See go.nature.com/xwf7yc for more.

Clinical publishing

The results of some US clinical trials are not being published completely enough, or even at all, according to an analysis published on 3 December (C. Riveros *et al.* *PLoS Med.* **10**, e1001566; 2013). A random selection of nearly 600 trials with results posted on the website ClinicalTrials.gov found that

ESA/NASA/SOHO/LASCO

50% had not had their results published in a paper. For trials that had been published in articles, data on negative side effects, adverse events and treatment efficacy were more clearly and completely reported on the website than in the paper. See go.nature.com/gqjbbk for more.

GM study retracted

Bowing to scientists' near-universal scorn, *Food and Chemical Toxicology* has retracted a controversial paper (G.-E. Séralini *et al.* *Food Chem. Toxicol.* **50**, 4221–4231; 2012) that claimed that Monsanto's genetically modified (GM) maize (corn) causes serious disease in rats. The authors had earlier refused to withdraw it. The paper showed “no evidence of fraud or intentional misrepresentation of the data”, said a 28 November statement from publishers Elsevier, but the small number and type of animals used in the study mean that “no definitive conclusions can be reached”. See go.nature.com/fk2auz for more.

China Moon rover

China has launched its first mission to land a rover on the Moon. The Chang'e 3 probe carrying the rover, which will survey lunar geology, blasted off on 2 December from the Xichang Satellite Launch



Center in southwest China (pictured). The spacecraft is designed to soft-land (rather than crash-land) on the Moon (see *Nature* **503**, 445–446; 2013), and is expected to touch down in mid-December. If the mission succeeds, China will become the third country behind the United States and the former Soviet Union to achieve a soft landing on the Moon.

Dance retraction

An eight-year-old *Nature* paper that reported a strong correlation between the symmetry of Jamaican dancers and their dancing ability was retracted on 27 November (W. M. Brown *et al.* *Nature* <http://doi.org/p9m>; 2013). No reason is given in the retraction notice, but in April co-author Robert Trivers, an evolutionary biologist at Rutgers University in New Brunswick, New Jersey, publicly released a university investigation report alleging that then-postdoc and

co-author William Brown had faked data for the paper (see *Nature* **497**, 170–171; 2013). Brown disputed the findings. Trivers has sought since 2008 to retract the paper, which implied that dancing ability might have evolved under sexual selection in humans.

BUSINESS

CRISPR company

The CRISPR technique for editing genes — which has rapidly become influential because of its efficiency and precision — has now been spun out into a commercial venture. Editas Medicine, based in Cambridge, Massachusetts, announced its launch last week, with an initial US\$43-million investment by four US-based venture-capital firms. It hopes to develop therapies by using CRISPR and other techniques to edit disease-related genes. Founding members include neuroscientist Feng Zhang of

COMING UP

9–13 DECEMBER

The American Geophysical Union hosts a meeting in San Francisco, California, including discussion of latest findings from the Mars Curiosity rover. fallmeeting.agu.org/2013

11 DECEMBER

London hosts a G8 Dementia Summit, where scientists and politicians will discuss the state of funding and research, and set out a global action plan for tackling dementia. go.nature.com/rv7tjz

the Massachusetts Institute of Technology in Cambridge, geneticist George Church of Harvard University in Boston, Massachusetts, and biochemist Jennifer Doudna of the University of California, Berkeley. See go.nature.com/yzhkci for more.

PEOPLE

Integrity advocate

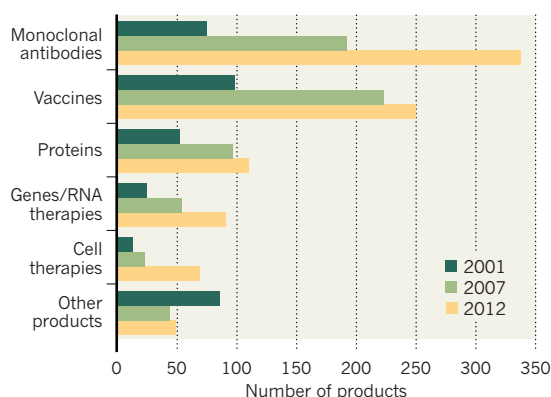
Scientist Francesca Grifo, an influential campaigner for reform of US government policies on scientific integrity, is moving to the Environmental Protection Agency (EPA) to oversee implementation of its own integrity policy. Grifo, previously at the Union of Concerned Scientists, an advocacy group based in Cambridge, Massachusetts, published an analysis in March comparing such policies across federal agencies. In it, she praised the EPA for clearly stating scientists' rights to express personal opinions with appropriate disclaimers, but noted gaps in online accessibility of the agency's key scientific-integrity documents.

TREND WATCH

Biologic drugs have grown increasingly dominant in the research and development efforts of pharmaceutical companies in the past decade — a marked shift from the industry's strong focus on small-molecule drugs in the 1990s. Biologic products, which include monoclonal antibodies, vaccines and cell therapies, made up 8% of total pharmaceutical sales in 2002 but 17% in 2012, according to a recent report by the Tufts Center for the Study of Drug Development in Boston, Massachusetts.

SHIFTING R&D STRATEGY

Between 2001 and 2012, the number of biologics in clinical development grew by 155%.



NEWS IN FOCUS

GENOMICS DNA from old bone unearths human-origins puzzle **p.16**

RESEARCH Italian police investigate allegations of scientific misconduct **p.18**

SCIENCE HISTORY That boring, boxy sequencer may be fit for a museum **p.20**



GLOBAL HEALTH Lessons learned from a million deaths in India **p.22**

KAVEH KAZEMI/GETTY



A woman undergoes chemotherapy in Tehran, where some cancer treatments are in short supply.

POLITICS

Iran hit by drug shortage

Sanctions cause increasing shortfall in medicines and vaccines.

BY DECLAN BUTLER

A tightening of already draconian international economic sanctions against Iran is causing serious shortages of certain drugs, vaccines and other key medical supplies in the country, medical researchers and public-health officials are warning.

The items, along with humanitarian goods such as food, are technically exempted from sanctions imposed by the United Nations, the United States and the European Union, which have strangled Iran's economy. But the sanctions' effects, for example on financial transactions, are causing shortages that are having a severe impact on hospitals, medical-research centres and the Iranian people, says

Ali Gorji, a neuroscientist at the University of Münster in Germany, and director of the Shefa Neuroscience Research Center in Tehran.

Exports of pharmaceuticals to Iran from the United States alone fell by half last year, from US\$31.1 million in 2011 to \$14.8 million. And Novartis, a pharmaceutical company based in Basel, Switzerland, says that the flow of life-saving products to Iran has been "severely affected, if not fully ceased".

Particularly affected are medicines and vaccines meant to treat and protect infants, as well as antibiotics and supplies for diagnostic equipment. As a result, lives are being put at risk, says Gorji.

A landmark deal to freeze the country's nuclear activity, reached last month by six

world powers and Iran, will see some sanctions relaxed, but Gorji and other experts are sceptical that it will have any immediate effect on alleviating the shortfall. Gorji adds that there is an urgent need for an independent outside assessment of Iran's medical shortages.

Gorji has been working to raise awareness of the problem since the beginning of the year, and in June organized a letter — signed by 70 scientists and physicians around the world — to UN secretary-general Ban Ki-moon asking him to address the situation. In his reply, Ban acknowledged that sanctions were having a detrimental effect on health, noting, for example, that "it is difficult, if not impossible, for importers to pay for medical supplies and equipment". He added that he was trying "to ensure that sanctions ►

► regimes have in place fair and clear procedures for granting humanitarian exemptions”.

Iran has a strong domestic drug and vaccine industry, producing 90% of its own medicines, but these are mostly generic drugs. The country has to import newer and more sophisticated drugs such as for cancer treatments, and imports of these specialist medicines have been hardest hit by sanctions, says Richard Garfield of Columbia University in New York, who studies the effects of conflicts and economic sanctions on public health. Iran's industry is also highly dependent on imports of raw materials, with difficult-to-source ingredients for more-complex drugs being most affected, he says.

According to Gorji, there are severe shortages of many drugs, including antibiotics, clofarabine for treating children with leukaemia and deferasirox for thalassaemia, a blood disorder common in Iran. He says that he witnessed the death of several children and adults as a result of drug shortages on a visit to Iran in early November. The routine child vaccine that protects against the bacterium *Haemophilus influenzae*, which causes severe pneumonia and meningitis in infants, is another casualty, adds Gorji. Also affected are parts and consumables for advanced medical technologies such as magnetic resonance imaging.

A major problem is the legal complexity of the sanctions — a maze of bans on certain transactions, individuals and organizations. Uncertainty as to what is and is not covered has led

many of the US, European and other companies that previously sold medical supplies to Iran to become reluctant to do so, says Garfield. Several companies, including Swiss financial firm Credit Suisse and London-based bank Standard Chartered, have incurred fines of hundreds of millions of dollars for falling foul of US sanctions against Iran, adding to the disincentive. Many firms are also nervous about being seen to be trading with an international pariah, Garfield adds.

Even when hospitals and research centres can find a supplier, tough sanctions on Iranian banks and foreign banks' dealings with Iran mean that they often cannot find a route to pay for medical supplies, says Seyed Hesamedin Madani, head of medical procurement for the Red Crescent Society of Iran in Tehran.

To make matters worse, sanctions affecting exports of Iranian oil have cut off the country's main supply of hard currency, says Madani, and the value of Iran's currency, the rial, has fallen by half against the dollar in the past 14 months, drastically increasing the price of medicines.

But the situation may be improving. In response to concerns from suppliers and banks about the difficulties in enacting exemptions for medical supplies, the US Department of

the Treasury's Office of Foreign Assets Control (OFAC) issued new guidance in July. The guidelines aim to reassure US and non-US medical suppliers and banks that exporting medicines and medical devices is 'broadly authorized', provided that it does not involve Iranian organizations proscribed under sanctions.

The guidance also expanded the list of medical supplies that can be exported without OFAC approval. Garfield says that OFAC's action is a step in the right direction, but thinks that without more proactive measures, the weight of disincentives for firms to engage with Iran means that little is likely to change.

Last month's interim agreement between Iran and the group of countries known as the P5+1 — the United States, the United Kingdom, France, Russia and China, plus Germany, and facilitated by the European Union — has also raised hopes. It puts Iran's nuclear programme on hold for six months, thus stalling any development of a nuclear weapon, and will see sanctions eased slightly in return.

The deal also includes a proactive provision to establish a financial channel to facilitate humanitarian trade with Iran. This would involve designating foreign and Iranian banks that are authorized for this purpose, and in principle would provide an official route for suppliers and Iranian hospitals and medical centres to carry out transactions. Garfield welcomes the channel as long overdue, but he warns that making it work is another matter. ■

Many firms are also nervous about being seen to be trading with an international pariah.

GENOMICS

Hominin DNA baffles experts

Analysis of oldest sequence from a human ancestor suggests link to mystery population.

BY EWEN CALLAWAY

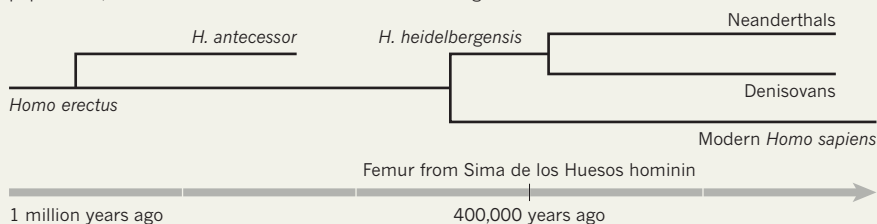
Another ancient genome, another mystery. DNA gleaned from a 400,000-year-old femur from Spain has revealed an unexpected link between Europe's hominin inhabitants of the time and a cryptic population, the Denisovans, who are known to have lived much more recently in southwestern Siberia.

The DNA, which represents the oldest hominin sequence yet published, has left researchers baffled because most of them believed that the bones would be more closely linked to Neanderthals than to Denisovans. “That's not what I would have expected; that's not what anyone would have expected,” says Chris Stringer, a palaeoanthropologist at London's Natural History Museum who was not involved in sequencing the femur DNA.

The fossil was excavated in the 1990s from

FAMILY MYSTERY

The mitochondrial genome of a 400,000-year-old femur has an unexpected link with a group of hominins called Denisovans. One interpretation is that this could be the result of interbreeding between more ancient populations, such as *Homo antecessor* and *Homo heidelbergensis*.



a deep cave in a well-studied site in northern Spain called Sima de los Huesos ('pit of bones'). This femur and the remains of more than two dozen other hominins found at the site have previously been attributed either to early forms of Neanderthals, who lived in Europe until about 30,000 years ago, or to *Homo*

heidelbergensis, a loosely defined hominin population that gave rise to Neanderthals in Europe and possibly humans in Africa.

But a closer link to Neanderthals than to Denisovans was not what was discovered by the team led by Svante Pääbo, a molecular geneticist at the Max Planck Institute for Evolutionary

Anthropology in Leipzig, Germany.

The team sequenced most of the femur's mitochondrial genome, which is made up of DNA from the cell's energy-producing structures and passed down the maternal line. The resulting phylogenetic analysis — which shows branches in evolutionary history — placed the DNA closer to that of Denisovans than to Neanderthals or modern humans. “This really raises more questions than it answers,” Pääbo says.

The team's finding, published online in *Nature* this week (M. Meyer *et al.* *Nature* <http://dx.doi.org/10.1038/nature12788>; 2013), does not necessarily mean that the Sima de los Huesos hominins are more closely related to the Denisovans, a population that lived thousands of kilometres away and hundreds of thousands of years later, than to nearby Neanderthals. This is because the mitochondrial genome tells the history of just an individual's mother, and her mother, and so on.

Nuclear DNA, by contrast, contains material from both parents (and all of their ancestors) and typically provides a more accurate overview of a population's history. But this was not available from the femur.

With that caveat in mind, researchers interested in human evolution are scrambling to explain the surprising link, and everyone seems to have their own ideas.

Pääbo notes that previously published full nuclear genomes of Neanderthals and Denisovans suggest that the two had a common ancestor that lived up to 700,000 years ago. He suggests that the Sima de los Huesos hominins could represent a founder population that once lived all over Eurasia and gave rise to the two groups. Both may have then carried the mitochondrial sequence seen in the caves. But these mitochondrial lineages go extinct whenever a female does not give birth to a daughter, so the Neanderthals could have simply lost that sequence while it lived on in Denisovan women.

“I've got my own twist on it,” says Stringer, who has previously argued that the Sima de los Huesos hominins are indeed early Neanderthals (C. Stringer *Evol. Anthropol.* **21**, 101–107; 2012). He thinks that the newly decoded mitochondrial genome



A dig at the Sima de los Huesos cave in Spain, the site of ancient hominin fossils.

may have come from another distinct group of hominins. Not far from the caves, researchers have discovered hominin bones from about 800,000 years ago that have been attributed to an archaic hominin called *Homo antecessor*, thought to be a European descendant of *Homo erectus*. Stringer proposes that this species interbred with a population that was ancestral to both Denisovans and Sima de los Huesos hominins, introducing the newly decoded mitochondrial lineage to both populations (see ‘Family mystery’).

This scenario, Stringer says, explains another oddity thrown up by the sequencing of ancient hominin DNA. As part of a widely discussed and soon-to-be-released analysis of high-quality Denisovan and Neanderthal nuclear genomes, Pääbo's team suggests that Denisovans seem to have interbred with a mysterious hominin group (see *Nature* <http://doi.org/p9t>; 2013).

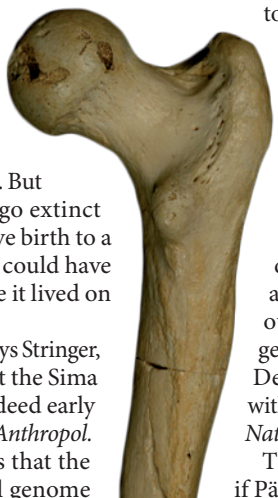
The situation will become clearer if Pääbo's team can eke nuclear DNA

out of the bones from the Sima de los Huesos hominins, which his team hopes to achieve within a year or so.

Obtaining such sequences will not be simple, because nuclear DNA is present in bone at much lower levels than mitochondrial DNA. And even obtaining the partial mitochondrial genome was not easy: the team had to grind up almost two grams of bone and relied on various technical and computational methods to sequence the contaminated and damaged DNA and to arrange it into a genome. To make sure that they had identified genuine ancient sequences, they analysed only very short DNA strands that contained chemical modifications characteristic of ancient DNA.

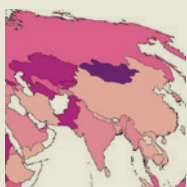
Clive Finlayson, an archaeologist at the Gibraltar Museum, calls the latest paper “sobering and refreshing”, and says that too many ideas about human evolution have been derived from limited samples and preconceived ideas. “The genetics, to me, don't lie,” he adds.

Even Pääbo admits that he was befuddled by his team's latest discovery. “My hope is, of course, eventually we will not bring turmoil but clarity to this world,” he says. ■



MORE ONLINE

TOP STORY

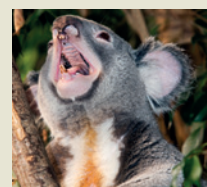


China and India drive global transition to meat-rich diets
go.nature.com/qhlnob

MORE NEWS

- Newlyweds' gut feelings offer prediction of marital happiness
go.nature.com/lj44xs
- Mice inherit pain associations epigenetically
go.nature.com/lf41np
- Camera takes pictures one photon at a time
go.nature.com/ytke4e

SOUNDBITE OF THE WEEK



Novel vocal organ explains koalas' deep bellows
go.nature.com/wehsuj

SCIENTIFIC MISCONDUCT

Image search triggers Italian police probe

Method for checking global literature leads to investigation of cancer researcher.

BY ALISON ABBOTT

As the editor-in-chief of two journals, cell biologist Gerry Melino is used to having an inbox full of concerns, of varying levels of seriousness, about data in published papers. But an e-mail he received criticizing a 2006 paper in one of his journals, *Cell Death and Disease*, surprised him with its unusual professionalism and precision. He fast-tracked an independent investigation into the paper — and this week the journal will publish a retraction by the authors, led by prominent Italian cancer researcher Alfredo Fusco. The paper had been found to contain “inappropriately duplicated” images.

The e-mail criticizing Fusco's paper was sent in August by Enrico Bucci, who runs a small biomedical start-up company in northern Italy offering publication of meta-analysis services. He had detected the gel anomalies while conducting a global search to exclude contaminated literature from his database. The search has revealed anomalous images in around one-quarter of biomedical papers examined so far.

Fusco, a professor at the University of Naples and an associate member of the Accademia dei Lincei, Italy's prestigious national academy, is now under investigation by the police and by his university.

Although Fusco is the first person to be investigated as a result of Bucci's work, other scientists may now find their work under scrutiny. The affair has also revealed the absence of a system for investigating allegations of misconduct in Italy's universities.

The story began in 2008, when Bucci, a molecular biologist, founded BioDigitalValley in Pont Saint Martin, Italy. Its services include pulling out all published images of gel-electrophoresis analysis — which separates and identifies large molecules such as proteins and sequences of RNA — that are relevant to a particular disease or tissue.

Bucci and his team created a database hosting all accessible biomedical papers published since 2000. But cleaning it of scientific contamination was not the quick job he had imagined. First he removed retracted papers; then he created a network of scientists who had been co-authors at least three times

with authors of the retractions.

The list ran to more than one million, so he looked only at Italian scientists. Using in-house software that could isolate images of, for example, gels, and check them for simple features such as possibly duplicated portions, he ran an automatic check of all the papers the Italian researchers had published. He focused on highly cited researchers for whom the



The University of Naples is carrying out an internal inquiry following allegations against one of its researchers.

automatic check had revealed multiple papers with anomalous images. Fusco, a specialist in cancer genetics, topped the list with eight papers.

Bucci's team then checked all of Fusco's research papers in greater detail using gel-checking software to identify various features, such as reused gel images or markings that suggested cut-and-pasted images. Out of around 300 papers on which Fusco was first or last author, the team found 53 containing gels with potential irregularities, including one from as far back as 1985.

After discovering that there was no academic organization in Italy that dealt with such findings, in February 2012 Bucci contacted the Milan police.

He says that his team is now working on tens of other cases of Italian scientists who scored highly in his automated check and collaboration network. “The value of my method, with its scientists' network element, is that the first quick check allows you to decide which authors to spend detailed time on,” says Bucci.

Now midway through the analysis, he

estimates that around one-quarter of the thousands of papers featuring gels that he has analysed so far potentially breached widely accepted guidelines on reproducing gel images. And around 10% seem to include very obvious breaches, such as cutting and pasting of gel bands. Some journals were more affected than others, he says. Those with a high impact factor tended to be slightly less affected. He plans to publish his results.

The public prosecutors must decide by April whether to bring charges against Fusco. A source close to the police investigation confirms that around 60 papers by Fusco are being studied, including a 2007 paper in the *Journal of Clinical Investigation* that was retracted by the editors in November. In an e-mail to *Nature*, Fusco declined to comment until the investigations are complete.

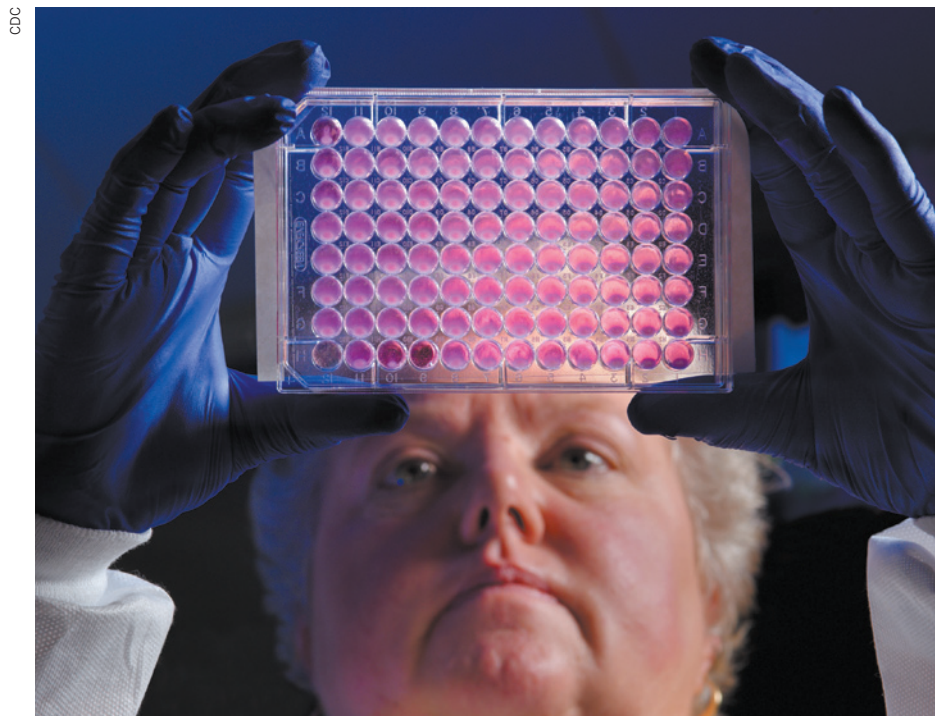
When details of the inquiry were leaked to an Italian newspaper in October, the rector of Fusco's university set up a three-man internal committee headed by Roberto Di Lauro, vice-rector for research at the University of Naples, that is expected to report by the end of the year. Di Lauro has co-authored nine papers with Fusco, but says that he will resign from the committee if any of these papers feature in the investigation.

Because universities in Italy have no established guidelines for handling such allegations, the committee is working out its operating procedures as it goes along, says Di Lauro. “We are taking the opportunity to make a report to our rectors on how universities should handle such cases,” he adds.

Such a report would be welcomed. “Italy is far behind other European countries in this,” says Nicole Föger of the Austrian Agency for Research Integrity in Vienna, and chair of the European Network of Research Integrity Offices. “But we have noticed that a country can catch up quickly after a scandal.”

“We editors are bombarded by low-quality allegations to the point that we can't cope,” says Melino, who has joint appointments at the MRC Toxicology Unit at the University of Leicester, UK, and the University of Rome Tor Vergata. “But Bucci's systematic approach to addressing the genuine problem of literature contamination is actually very helpful.” ■

CONTRASTO/EVEVINE



Tests for antibiotic resistance in bacteria require regular updates to remain effective as microbes evolve.

MICROBIOLOGY

Bacteria evade detection net

Slothful response from regulators and manufacturers means antibiotic resistance is missed.

BY SARAH ZHANG

Bacteria that are resistant to almost all antibiotics are dreaded by physicians and patients alike. Finding such microbes in a hospital is bad enough, but failing to detect them can lead to something much worse: an outbreak.

That is why patients with bacterial infections are typically screened using automated lab tests that uncover the pathogen's identity and its susceptibility to various antibiotics. Because mechanisms of antibiotic resistance can pass rapidly between bacteria, especially in hospitals, the US Food and Drug Administration (FDA) periodically updates the rules for the machines used to perform these tests.

But there is a problem. The latest FDA rules have still not been fully adopted in tests for carbapenem-resistant Enterobacteriaceae, or CRE. This particularly worrisome group of pathogens can overcome carbapenem

antibiotics — among the last lines of defence against resistant infections.

Each time the FDA updates its susceptibility criteria, manufacturers must upgrade the screening machines and resubmit them to the FDA for approval. Currently, none of the three manufacturers of such devices has incorporated the latest criteria for all three major types of carbapenem antibiotic. Most hospitals continue to use machines with outdated software dozens of times a day. That means CRE cases are probably going undetected, says Janet Hindler, a clinical microbiologist at the University of California, Los Angeles. "It is sad this is a man-made problem," she says.

Identifying resistant bacteria matters both for surveillance and for individual patients, whose antibiotic treatments are tailored to the susceptibility of the infection, as determined by screening. The process uses a plastic plate with small

wells containing up to a dozen antibiotics at different concentrations. The machine analyses bacterial growth in each well for 4–16 hours to determine the minimum amount of antibiotic needed to kill the pathogen. This antibiotic concentration, under the FDA rules, classifies the bacteria as susceptible, intermediate or resistant.

The resistant category keeps getting larger. One reason is that a better understanding of how drugs work in people — as opposed to in a Petri dish — shows that the threshold for antibiotic resistance in the body is lower than once thought. Another reason is that many types of resistant bacteria have emerged since various antibiotics were first approved. The FDA's drug-evaluation section is currently updating the classifications for a backlog of 200 antibiotics.

The process of updating the US criteria for antibiotics starts with work by the Clinical and Laboratory Standards Institute (CLSI), a non-profit organization in Wayne, Pennsylvania, that collects drug data and helps to set the guidelines. After a decade of work on carbapenems, it eventually published revised classifications in 2010, including a larger resistant category for CREs. But the FDA did not start to implement its own categorization rules until 2012. The screening devices themselves are still out of date, so a bacterium considered resistant under the latest FDA criteria could still be reported as susceptible or intermediate.

Europe has removed at least one hurdle to getting its own up-to-date resistance categories. The expert body that recommends the classifications, the European Committee on Antimicrobial Susceptibility Testing, also sets the criteria for the European Medicines Authority. And companies can make small updates to their devices without having to reapply to regulatory bodies for approval, says David Livermore, a microbiologist at the University of East Anglia in Norwich, UK. In the United States, he says, "making updates is a rather tortuous business, particularly if the CLSI has changed its breakpoints and the FDA hasn't".

Manufacturers concur with Livermore. Before it will clear an updated device, the FDA sometimes requires further studies that can take two to three years to conduct and cost tens of thousands of dollars, says Bill Brasso, president of the Susceptibility Testing Manufacturer's Association and a scientist at one of the manufacturers, BD Diagnostics Systems in Sparks, Maryland. It can then take up to a year to get FDA clearance, he adds.

But Sally Hojvat, director for the division of microbiology devices at the FDA, says that data from earlier studies will often suffice. Even when new studies are required, she says, they involve only the analysis of bacteria, not large clinical trials. Sometimes the process drags on because manufacturers do not submit all the information the FDA needs, she adds.

Neither manufacturers nor the FDA has pushed for timely updates, says Amy Leber, director of the microbiology lab ▶

"Making updates is a rather tortuous business."

► at the Nationwide Children's Hospital in Columbus, Ohio. "The manufacturers point their fingers at the FDA, and the FDA points its finger at the manufacturers," she says.

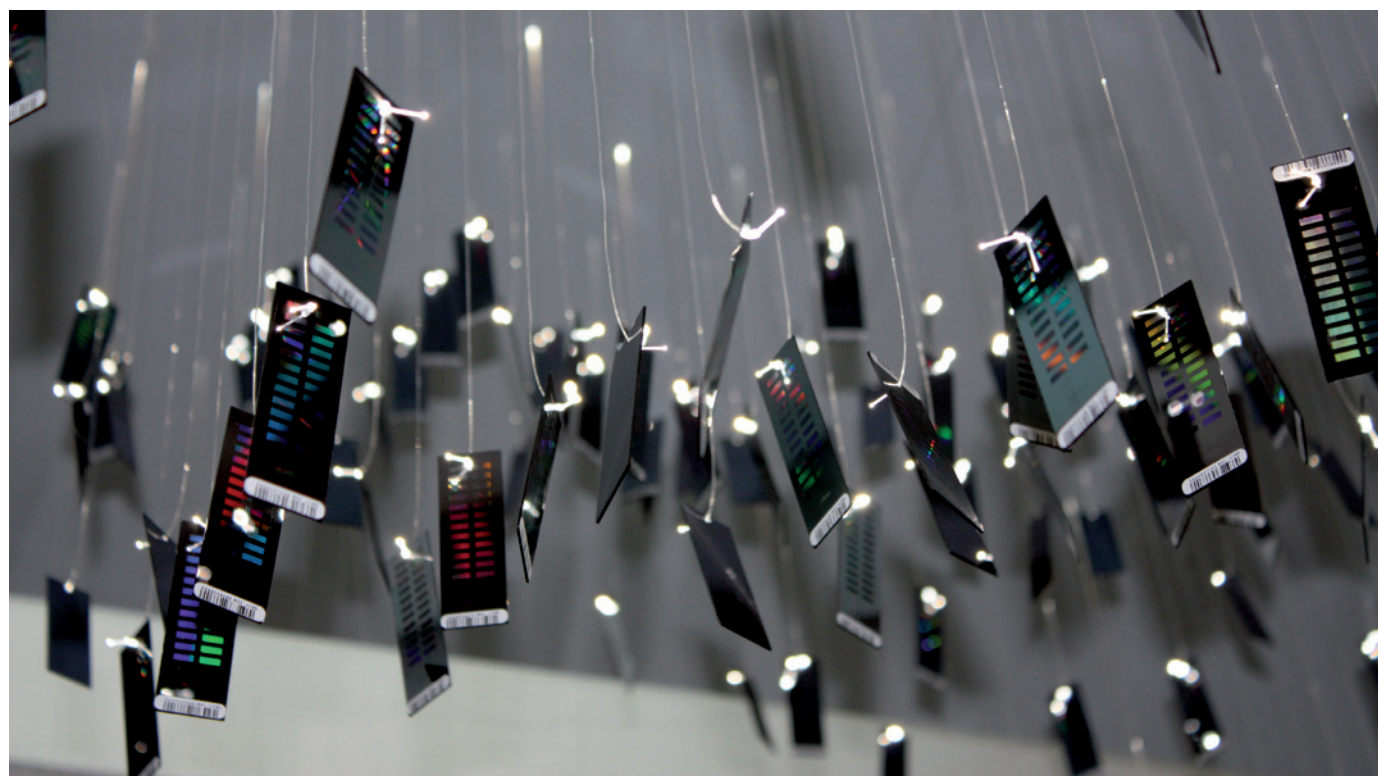
Leber found her own way around the problem. Last year, her lab colleagues hacked into her hospital's machine to upgrade the software that governs resistance categories for carbapenems. Earlier this year the machine

identified a case of CRE. The patient was swiftly isolated and an outbreak averted.

Leber worries about smaller clinics that do not have advanced screening equipment, let alone the time and money to update their devices. Meanwhile, CRE cases are spreading out of urban centres. Since the microbes were first detected in North Carolina in 2001, they can now be found in nearly every US state.

Jean Patel, deputy director of the office of antimicrobial resistance at the Centers for Disease Control and Prevention in Atlanta, Georgia, wants screening devices updated faster so that her agency can conduct better surveillance. "It has been a little frustrating to watch how long this has taken." ■

Additional reporting by Elizabeth Gibney.



MEDICAL MUSEION/UNIV. COPENHAGEN

Illuminated microarrays formed part of a display at a medical museum in Copenhagen in 2011.

SCIENCE HISTORY

Museums hunt for relics from genomics' early days

Collectors band together to salvage cast-off equipment.

BY HEIDI LEDFORD

In a former envelope factory sits a boxy grey and blue machine the size of an oven — the tenth acquisition this year for a Massachusetts science museum. It is a colony picker, a robotic arm that plucks bacteria from Petri dishes and drops them into a tray with 96 wells, from which DNA is extracted, amplified and sequenced.

At the start of this century, the device

powered genomics research at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. Now retired, it rests in the warehouse of the Massachusetts Institute of Technology (MIT) Museum, sandwiched between a sewing machine from around the 1920s and an analog computer from the 1950s.

Genomics researchers

► **NATURE.COM**
For a slideshow of equipment sought by curators, see: go.nature.com/o5ruqh

might find it hard to believe that museum patrons would be interested in paying homage to a piece of lab equipment as pedestrian as a colony picker. Curators of science and technology museums say that is exactly the problem.

"Very few scientists have any idea that they should preserve these things," says Thomas Söderqvist, director of the Medical Museion in Copenhagen. "They just throw them out."

But in the past two years, a confederation of about a dozen science museums has

coalesced around the need to preserve relics of the genomics revolution, in an effort known as the Museum Genomics Initiative. It was born of a concern that, in a time of shrinking museum budgets, the collection of scientific artefacts was not keeping pace with innovation. This shortcoming has been felt across disciplines, says Simon Chaplin, head of the Wellcome Collection's library of biomedical history in London, which has also joined the initiative. But he says that an effort focused on genomics makes sense because of the field's importance for medicine, appeal to the public and rapid growth since the late 1990s. "There's a real risk that if we don't act quickly, the material legacy of genomics will be lost," he says.

Such was nearly the fate of one of the colony picker's neighbours, a machine with a conveyor belt running along its top. Its job was once to shuttle a colony picker's 96-well trays between stations (each named after a stop on the subway line that runs through Cambridge) to prepare samples for sequencing. John Durant, director of the MIT Museum in Cambridge, came across it about a year ago while rummaging through a storage facility at the nearby Broad Institute.

The machine was slated for disposal, even though it had been used during the peak of the frenzy to sequence the first human genome. "We looked at this thing and said immediately, 'We'll have it,'" says Durant.

He likens his job to that of a contemporary-art collector: he has to predict what items will hold value decades from now. Scientific advisers help curators in this assessment. Robert Bud, chief curator of science and medicine at the century-old Science Museum in London — home of 'Baby Blue', a prototype machine for running the polymerase chain reaction to amplify bits of DNA — says that the Museum Genomics Initiative aims to help museums to prioritize and consolidate their efforts by creating a list of pieces recommended for acquisition. Bud declines to name all the items he would put at the top of his own wish list, however: "The moment I say something, it acquires value."

Luckily, unlike contemporary art, cast-off lab equipment rarely comes at a high price. Instead, the cost lies in storage, particularly for large pieces. And if museums want to keep the machines in working order, finding the right consulting technicians and spare parts can be costly, says Heather Erickson, president of the Life Sciences Foundation in San Francisco, California, a non-profit organization dedicated to preserving historical information about biotechnology. (A colony



'Baby Blue', an early DNA amplifier.

picker is striking when its robotic arm is working, but little more than a box when it is not.)

Sexing up the visual appeal of the artefacts is another challenge, says Söderqvist. Over the past 50 years, as electronics became miniaturized and manufacturing was standardized, the beautifully customized machines of old gave way to uninspiring grey boxes. "We are working with more and more abstract objects," he says. "Does a DNA sequencer look any different from your dishwasher?"

Söderqvist sees his role in the initiative as providing some visual pizzazz to these DNA 'dishwashers'. In 2011, he helped to create an exhibition of microarrays (slides coated with 20,000 unique DNA fragments) used in a diabetes experiment. His museum drilled holes in about 600 arrays, and strung them from the ceiling, illuminating them with fibre optics.

Some items have more obvious appeal and are objects of acute desire for curators. Durant gets a dreamy look when he discusses the display that was hung in the reception area of the Wellcome Trust Sanger Institute in Cambridge, UK, in the mid-1990s during the Human Genome Project. A digital ticker scrolled through the DNA letters that had come up in the previous day's sequencing — and the rate at which the As, Ts, Cs and Gs flew by underscored not only advances in sequencing technology, but also the institute's mission to make those sequences publicly available.

The Sanger still has the sign, and sometimes trots it out for visiting school groups, but it no longer greets visitors in reception because the system cannot keep up with modern sequencing speeds. Bud says that his museum would like to acquire it.

Also on Bud's agenda is a sequencing machine from UK company Oxford Nanopore Technologies. The machines, some of which can sequence the human genome in 15 minutes, are not yet relics; they have not been commercially released and labs around the world are queuing up to access the first batch (see *Nature* <http://doi.org/p8j>; 2012). "It's going to be among the hardest to acquire," says Bud. "But we've been around a hundred years. We'll wait." ■

CORRECTION

The News story 'China aims for the Moon' (*Nature* **503**, 445–446; 2013) should have said that Chang'e-3 will deploy the first near-ultraviolet telescope on the Moon (*Apollo 16* used a far-ultraviolet telescope).

ONE MILLION DEATHS

WHAT RESEARCHERS ARE LEARNING FROM AN UNPRECEDENTED SURVEY OF MORTALITY IN INDIA.

BY ERICA WESTLY

In 1975, when Prabhat Jha was growing up in Canada, his family received a report from India that his grandfather had died; the cause was unclear. Like many people living in rural India, Jha's grandfather had died at home, without having visited a hospital. Jha's mother was desperate for more information, so she returned to her home village to talk to locals. Years later, when Jha was at medical school, he reviewed his mother's notes and realized that his grandfather had probably died of a stroke. Now Jha, an epidemiologist at the University of Toronto, is nearing the end of an ambitious public-health programme to document death in India using similar 'verbal autopsy' strategies.

The Million Death Study (MDS) involves biannual in-person surveys of more than 1 million households across India. The study covers the period from 1997 to the end of 2013, and will document roughly 1 million deaths. Jha and his colleagues have coded about 450,000 so far, and have deciphered several compelling trends that are starting to lead to policy changes, such as stronger warning labels on tobacco.

Public-health experts need mortality figures to monitor disease and assess interventions, but quality mortality data are scarce in most developing countries. Seventy-five per cent of the 60 million people who die each year around the globe are in low- and middle-income countries such as India, where cause of death is often misclassified or unreported. Groups such as the World Health Organization (WHO) typically base mortality estimates on hospital data, but in many developing countries most people die outside hospitals.

As global health researchers increasingly turn to indirect computer models, many applaud the MDS's low-tech, on-the-ground approach and see it as a model for assessing true health burdens in the developing world. "For countries like India, there will almost certainly continue to be a role for verbal autopsy," said Colin Mathers, coordinator of mortality and burden of disease at the WHO. "It's a crucial source of information."

HOW THEY GATHER THE DATA

The Million Death Study (MDS) involved two phases, 1997–2003 and 2004–2013, each of which surveyed a different selection of more than 1 million homes.

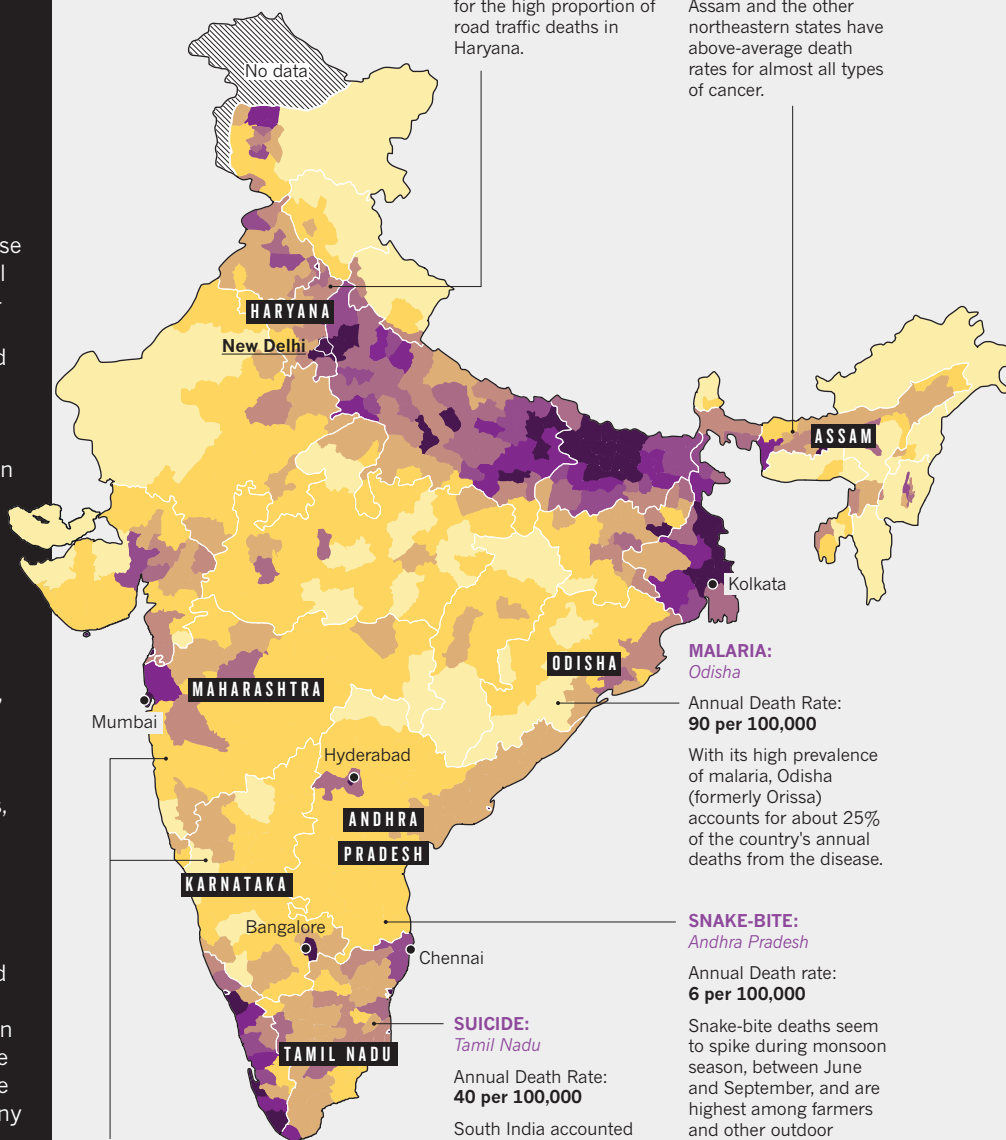
800-900

government surveyors visit the homes every six months.



MAPPING MORTALITY

The project data show that cause of death is influenced by geography. Knowing which threats are greatest in which states informs policies and future studies.



ROAD-TRAFFIC INJURY: Haryana

Annual Death Rate: 30 per 100,000

High-density trucking routes may be to blame for the high proportion of road traffic deaths in Haryana.

CANCER: Northeastern States, including Assam

Annual Death Rate: 65 per 100,000

For reasons not yet clear, Assam and the other northeastern states have above-average death rates for almost all types of cancer.

MALARIA: Odisha

Annual Death Rate: 90 per 100,000

With its high prevalence of malaria, Odisha (formerly Orissa) accounts for about 25% of the country's annual deaths from the disease.

SNAKE-BITE: Andhra Pradesh

Annual Death rate: 6 per 100,000

Snake-bite deaths seem to spike during monsoon season, between June and September, and are highest among farmers and other outdoor labourers.

SUICIDE: Tamil Nadu

Annual Death Rate: 40 per 100,000

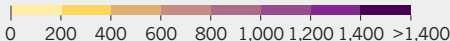
South India accounted for more than 40% of India's suicides. The area has high education levels and unemployment, both considered risk factors for suicide in India.

HIV: Maharashtra/Karnataka

Annual Death Rate: 56 per 100,000

Rural areas around Mumbai, the capital of Maharashtra, have the highest concentration of HIV-related deaths in India.

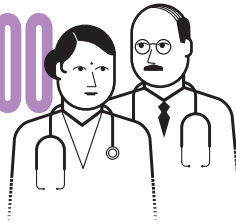
POPULATION DENSITY (people per km²)



DESIGN BY JASIEK KRZYSZTOFIAK/NATURE

50,000–58,000 TWO

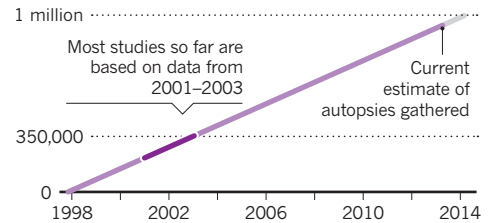
verbal autopsies are collected each year.



trained doctors from a pool of 300 assign a cause of death on the basis of each autopsy.

REACHING 1 MILLION

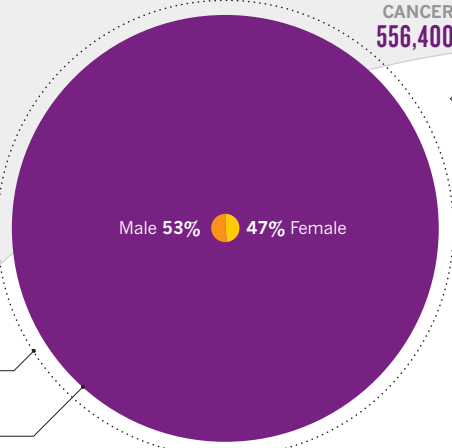
Owing to some delays related to the 2011 national census, the researchers will not have data on all 1 million deaths for a few more years.



PRECISION AND CONTROVERSY

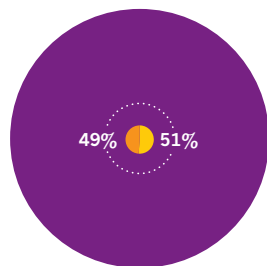
Studies based on the MDS's data help to provide a detailed picture of death in India, particularly for adolescents and adults living in rural areas. The findings include some surprising deviations from World Health Organization estimates.

WHO estimate
MDS figure



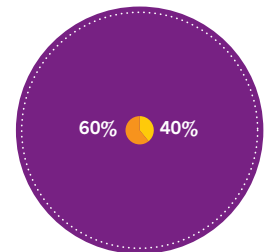
CANCER

Rural areas had a higher incidence of deaths from infection-related cancers, such as stomach and cervical cancer. Cervical cancer was the top killer for women, suggesting that interventions such as screening and HPV vaccination should be increased.



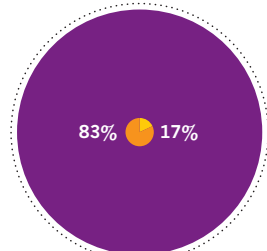
MALARIA

Controversially, the MDS's estimate for malaria deaths was much higher than the WHO's. The MDS found that about 58% of malaria deaths occurred in people aged 15–69.



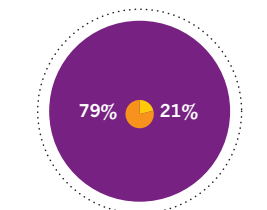
SUICIDE

The suicide rate in Indian women of 15 years and older is more than 2.5 times that for women of the same age in high-income countries. Poisoning — often with pesticides — is the most common method for both sexes.



ROAD-TRAFFIC INJURY

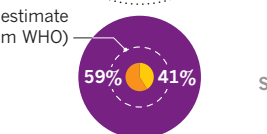
Two-wheel riders and pedestrians accounted for some 65% of traffic deaths. And about 60% involved head injury, suggesting interventions such as increased helmet use, lower speed limits and protected walkways for pedestrians.



HIV

The prevalence of HIV in India is relatively low, but the country has the third-largest number of people living with HIV in the world owing to its large population.

Previous estimate (not from WHO)

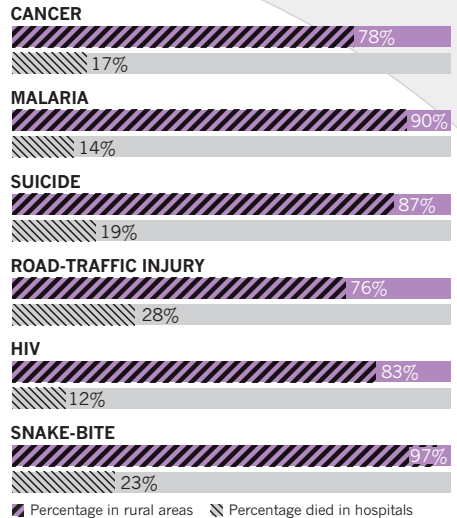


SNAKE-BITE

Estimates based on hospital data may miss many such deaths because three-quarters of them occur outside hospitals. Community education and increased distribution of anti-venom to rural areas are the main interventions.

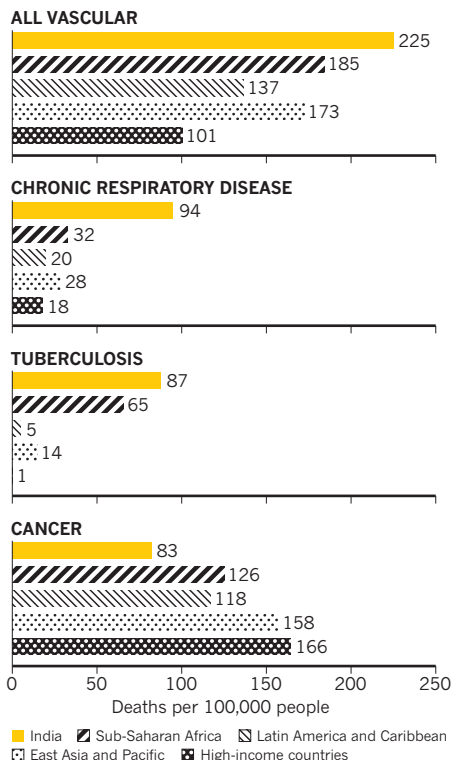
THE HIDDEN DEATHS

Most deaths are occurring outside of the hospital setting and in rural areas where they are often not registered. The MDS is starting to expose the hidden trends.



TOP CAUSES OF DEATH

The MDS determined that the four most significant causes of death for Indians aged 30–69 are vascular disease, chronic respiratory disease, tuberculosis and cancer. Some of these burdens look very different in other regions of the world.



SOURCE: P. JHA ET AL. PLOS MED. 3, E18 (2006); R. DIKSHIT ET AL. LANCET 379, 1807–1816 (2012); N. DHINGRA ET AL. LANCET 376, 1768–1774 (2010); V. PATEL ET AL. LANCET 379, 2343–2351 (2012); M. HSIAO ET AL. BRIT. MED. J. OPEN 3, E002621 (2013); P. JHA ET AL. BRIT. MED. J. 340, C621 (2010); B. MOHAPATRA ET AL. PLOS NEGL. TROP. DIS. 5, E1018 (2011).



NORTH AMERICA'S BROKEN HEART

A billion years ago, a huge rift nearly cleaved North America down the middle. And then it failed. Researchers may be getting close to finding out why.

BY JESSICA MARSHALL

On a bright October Saturday, the trees have reached their full autumn blaze at Interstate State Park on the border between Minnesota and Wisconsin in the US heartland. Crowds of people gawking at leaves thread their way along paths that wind through bulwarks of dark basalt, leading to views of the St Croix River. Along one of the walkways, a photographer directs a young couple in coordinating grey shirts to lean against the rocks as she snaps a romantic portrait.

If the two sweethearts are looking to commemorate their everlasting

love, they should have picked a different backdrop. The fractured basalt that frames their faces is part of a great gash that opened up in the middle of North America and nearly split the continent 1.1 billion years ago — hardly a symbol of a happy union.

The volcanic rocks are remnants of what is called the Midcontinent Rift, and it is an enormous geological puzzle. Rifts are wounds in Earth's outer layer that can grow to eventually form new oceans. That is how the

Basalt cliffs along the St Croix River are remnants of rifting 1.1 billion years ago.

Atlantic Ocean got its start some 200 million years ago, and an active rift continues to widen that basin. But the Midcontinent Rift was different. It opened a 3,000-kilometre crack in North America and created a basin as big, perhaps, as the Red Sea — then the system shut down. The wound stopped growing and the continent remained intact.

“How that feature could just totally reorganize the crust of the Earth in the Lake Superior region and not manage to break the continent apart is fairly amazing,” says G. Randy Keller, a geophysicist at the University of Oklahoma in Norman and director of the Oklahoma Geological Survey. “It’s a spectacular failure.” And a forgotten one, too. The rift is mostly buried under thick sediments, which makes it hard to study. And it lies far from the continent’s attention-grabbing geological features, such as mountain belts and earthquake zones. “For a long time, the rift has been a very neglected thing,” says Peter Hollings, a geochemist at Lakehead University in Thunder Bay, Canada.

That is now changing. Geologists have started to flock to the region to explore the enormous deposits of ore minerals left by volcanic activity during the creation of the rift: one area in northern Minnesota, for example, is the largest untapped copper–nickel deposit in the world. Another source of interest has come from the US National Science Foundation’s EarthScope project and related programmes, which installed dozens of temporary seismometers across the rift to provide an unprecedented picture of Earth’s crust and upper mantle there.

Researchers are keen to test theories about why the rift began and failed — and to use the ancient wound to improve more general understanding of how plates move and break apart. What is more, because the lava flows in the rift are stuck in the middle of a continent, they have been left as they were 1 billion years ago, unmangled by the collisions that warp rocks at the edges of continents. The basalts therefore offer an unparalleled record of events on Earth at a time when the continental plates were assembling into a supercontinent dubbed Rodinia, not long after multicellular life evolved.

Among researchers, there is a sense that the rift’s time has come. “There’s a whole flood of interest on the part of geoscientists who really weren’t interested before,” says Keller.

LISTENING POSTS

Some 145 kilometres northeast of where the couple posed for its picture, Suzan van der Lee leans into her shovel, grey hair tucked under a bandana. About a metre below the forest floor, she uncovers a seismometer buried in a black plastic pipe. A geophysicist at Northwestern University in Evanston, Illinois, she is there with a graduate student, Emily Wolin, who squats in front of a laptop as she backs up data from the instrument.

The site is 50 kilometres south of Lake Superior, well off the main road amid a dense stand of aspen and oak saplings. Wolin selected the spot two-and-a-half years ago by touring the region, seeking places in or near the Midcontinent Rift that were far enough from roads to avoid vibrations from traffic.

Since then, Wolin has been monitoring the stations every six months. On her rounds, she has had to flee an angry dog, don skis after a late-spring snow and seek help from a bear hunter to rescue her car from mud. One of her stations was burned in a wildfire (it still worked, even though a cable had melted), and another recorded the vibrations of trees crashing down in a massive windstorm. At a different site, a hunter apparently used the solar panel that powered the instrument for target practice: Wolin found a bullet hole right through its centre.

Today the team is here to recover instruments that have weathered two winters while quietly logging seismic activity across the globe. The stations are part of an EarthScope accessory project known as SPREE, or the Superior Province Rifting EarthScope Experiment. The project aims

to fill in details about the Midcontinent Rift by installing extra stations — 82 in total — tracing and transecting it. The seismometers provided what amounts to medical scans of the top 1,000 kilometres of crust and mantle near the rift. Van der Lee hopes to use that to better understand what is down there, learn how deep the rift extends and perhaps gather some clues as to what caused it.

Even though the researchers were careful to site the seismometers in quiet spots, the data coming back contain an inevitable amount of noise. The instruments are so sensitive that they detect not only earthquakes all over the world, but also noise from oceans and all kinds of other seismic background activity. The challenge is to pick that apart at each station and extract a real signal. The team is in the thick of that now and it will be months before it has a fuller picture of the subsurface structure.

But the picture that has emerged so far has been intriguing: the data show a significant amount of variation along the rift. “All the things that we’re seeing suggest that we’re dealing with a very complex structure,” says van der Lee.

SECRETS OF THE DEEP

SPREE and other geophysical studies will be key to unlocking the rift’s deepest secrets, because most of the structure is hidden. From magnetic and gravity surveys of the region over the past

half-century, geophysicists have determined that the rift is shaped like a horseshoe, with two arms pointing south from Lake Superior (see ‘Breaking up is hard to do’). Seismic studies in the 1980s revealed that the rift’s basalt layers reach deep below ground, up to 30 kilometres below Lake Superior¹. All told, the rift produced between 1 million and 2 million cubic kilometres of basalt, making it one of the world’s largest deposits of that rock.

The sheer volume of erupted basalt has led many to suggest that the rift must have been fed by a mantle plume — a vertical stream of hot rock rising from the depths of the planet. Another widely held idea is that tectonic plates smashing into the continent from the east stopped the rift from growing.

Researchers have used techniques in geochemistry, geophysics and other fields to test these ideas, with conflicting results. “There are problems with all of the models,” says Hollings. “None of them really works perfectly.” Still, he adds: “All the new work that’s being done is allowing us to re-evaluate the models and look at different ways this could have happened.”

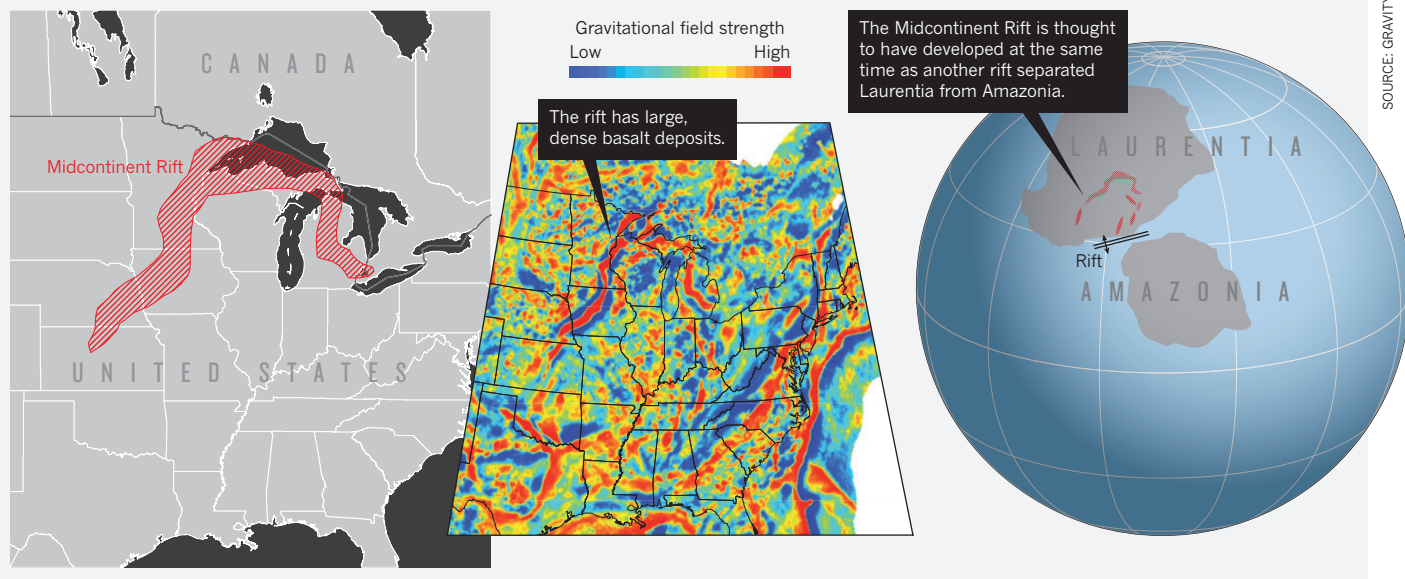
Some researchers are starting to see the rift as part of a much bigger puzzle. Geophysicist Carol Stein of the University of Illinois at Chicago discussed this idea at a meeting of the Geological Society of America (GSA) in Denver, Colorado, in October. Working with Keller and others, she suggests that the Midcontinent Rift was not isolated, but in fact was connected to other rifts that caused large changes in Earth’s tectonic plates at the time — all related to the assembly of Rodinia. The hypothesis is based on previous work² by Keller, who has proposed that gravity maps show the arms of the rift extending much farther to the south than was thought, towards the edge of Laurentia, the precursor to the North American continent. Other studies have suggested that Laurentia and Amazonia, the precursor to part of the South American continent, were in contact more than 1 billion years ago, and that they began to separate around the time that the Midcontinent Rift became active³. Stein proposes a three-armed rift system, formed of the Midcontinent Rift and the two arms that split Laurentia and Amazonia.

“In a lot of locations, when the continents break apart you seem to have three arms where one arm will fail and together the two arms make a new ocean,” she says. “It’s a lot less mysterious than we used to think even a year ago. We used to think of the Midcontinent Rift as this kind of weird feature that started and died in the middle of a continent.” But

“It’s a lot less mysterious than we used to think even a year ago.”

BREAKING UP IS HARD TO DO

Just over 1 billion years ago, North America started to split, opening a rift that filled with volcanic rocks. The wound healed, but left a horseshoe-shaped scar of dense rocks (left) that shows up on a gravity map of the central United States (centre). One theory proposes that the rifting was connected to the separation of Amazonia, now part of South America, from Laurentia, part of modern North America (right).



SOURCE: GRAVITY MAP: G. R. KELLER; GLOBE: C. STEIN

considered in connection with the rifts at the edge of Laurentia, it makes sense, she says. Stein likens the ancient system to what is happening on the eastern edge of Africa today. Two rift arms in the Gulf of Aden and the Red Sea are pushing the Arabian peninsula away from Africa, and another is starting within Africa. If that East African Rift fails to grow and eventually dies, it will end up looking like North America's Midcontinent Rift, she says.

Stein's hypothesis has piqued the interest of other experts. "It's a very reasonable idea," says Stephen Marshak, a geologist at the University of Illinois at Urbana-Champaign, although he feels that more testing must be done. He and others agree that understanding the Midcontinent Rift will provide insight into the East African Rift — what is driving it and how it might propagate in the future. "They are both informing each other," he adds.

HOT ROCKS

Apart from trying to understand the rift, researchers are also interested in using the feature as a window on the past. Protected in the stable centre of North America, the rift's lava flows have remained undisturbed for 1 billion years — a rarity for rocks that old. In some places, ripples that formed as the lava cooled are still visible on the basalt. "It's gorgeous," says Nicholas Swanson-Hysell, a geologist at the University of California, Berkeley. "How well these flows are preserved is pretty amazing. You could go to a lava field in Hawaii that erupted in 1950 and the surface would look similar to this 1.1-billion-year-old surface."

Just this kind of preservation is on display at Mamainse Point on the eastern shores of Lake Superior, where Swanson-Hysell has sampled 95 lava flows along 10 kilometres of shoreline. The individual flows, which range from a few metres to 20 metres thick, are part of a 4.5-kilometre-thick formation of rock created during the most active 15 million years of the rift's 30-million-year lifetime.

Magnetic grains in these rocks captured the orientation of Earth's magnetic field at the time the lava cooled. The readings from these minute frozen compasses can be used to track how Laurentia wandered around the globe during the span of the rifting. When Swanson-Hysell constructed a magnetic record from the Mamainse flows, he found signs that Laurentia may have been moving faster than any other plate is known to have travelled⁴. His latest estimates, presented at the GSA

meeting, put its velocity between 16 centimetres and 45 centimetres per year. For comparison, the next-fastest known plate movement is India's 18-centimetre-per-year rush towards Asia between 50 million and 60 million years ago. "This velocity is considered to be very fast and near the maximum rate possible for continental motion," says Swanson-Hysell. Most plates today move only around 4–9 centimetres per year. The range Swanson-Hysell has calculated for Laurentia is broad, but he aims to narrow it in the future.

Knowing how fast the continent moved gives researchers important information to help them to reconstruct the motion of all of Earth's landmasses at the time the rift formed. Swanson-Hysell says it is possible that the extraordinary velocities recorded there reflect more than just Laurentia's movement. Some of the motion could have been caused by a phenomenon called true polar wander, in which the whole crust and mantle rotate together around the core. This would happen if an extra-dense blob of material in the mantle was migrating towards the equator, taking the crust and mantle with it.

If there was true polar wander, it would be a sign of "something big happening in the interior of Earth," says Swanson-Hysell. Even if there was not, he adds, the high speed of Laurentia could provide insight into what was driving the motion of the tectonic plates at the time. The truth could be a combination of the two. To test this, Swanson-Hysell wants to construct similar records for sets of rocks on other continents. If they show the same fast motion, it would demonstrate that all the plates were moving together, pointing to true polar wander. But it is no easy task to find such well-preserved rocks from so long ago.

Back at Interstate State Park, it begins to drizzle, and the crowds head back to their cars. The raindrops darken the basalt, momentarily giving it the look of a fresh lava flow. Soon the site will empty and only the rocks will remain, full of history that geologists are just starting to unravel. ■

Jessica Marshall is a writer in St Paul, Minnesota.

1. Cannon, W. F. *et al. Tectonics* **8**, 305–332 (1989).
2. Adams, D. C. & Keller, G. R. *Can. J. Earth Sci.* **31**, 709–720 (1994).
3. Elming, S.-Å. *et al. Geophys. J. Int.* **178**, 106–122 (2009).
4. Swanson-Hysell, N. L., Maloof, A. C., Weiss, B. P. & Evans, D. A. D. *Nature Geosci.* **2**, 713–717 (2009).

COMMENT

PHYSICS Richard Feynman's lectures, still loved 50 years on **p.30**



EDUCATION Weighing the case for gene streaming in schools **p.32**

HEALTH Microbe-rich life with the Hadzabe hunter-gatherers of Tanzania **p.33**

OBITUARY Leonard Herzenberg, immunology technology pioneer, remembered **p.34**

BRIAN A. VIKANDER/COREIS



The Moon has a similar composition to the outer portions of Earth.

Lunar conspiracies

Current theories on the formation of the Moon owe too much to cosmic coincidences, says **Robin Canup**. She calls for better models and a mission to Venus.

The Moon is more than just a familiar sight in our skies. It dictates conditions on Earth. The Moon is large enough to stabilize our planet's rotation, holding Earth's polar axis steady to within a few degrees. Without it, the current Earth's tilt would vary chaotically by tens of degrees. Such large variations might not preclude life, but would lead to a vastly different climate.

Knowing how the Moon was made is central to understanding Earth and the formation of other planets. Since the 1980s, work on lunar origins has focused on the 'giant-impact' theory. This proposes that the

collision of another planet-sized body with the forming Earth generated a disk of debris that coalesced into the Moon. Such giant collisions were common in the Solar System during the final stages of Earth's formation 4.5 billion years ago.

But we still do not understand in detail how an impact could have produced our Earth and Moon. In the past few years, computer simulations, isotope analyses of rocks and data from lunar missions have raised the possibility

of new mechanisms to explain the observed characteristics of the Earth-Moon system.

The main challenge is to simultaneously account for the pair's dynamics — in particular, the total angular momentum contained in the Moon's orbit and Earth's 24-hour day — while also reconciling their many compositional similarities and few key differences. The collision of a large impactor with Earth can supply the needed angular momentum, but it also creates a disk of material derived largely from the impactor. If the infalling body had a different composition from Earth, as seems probable given that most objects in ▶

NATURE.COM
For more on the
Moon's origins, see:
go.nature.com/5foh6i

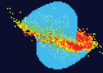
HOW TO MAKE A MOON

Simulations reveal how a giant collision between two similarly sized planets ('half-Earths') might explain why the Moon has a similar composition to Earth's mantle. The violent crash blended both planets to produce Earth and a disk of hot debris that coalesced to form the Moon (blue to red spans temperatures from below 2,000 kelvin to more than 6,400 kelvin).

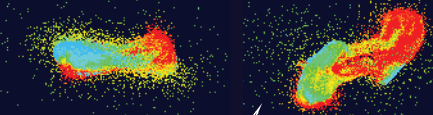
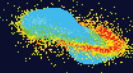
➔ FOR AN ANIMATED VERSION
SEE: go.nature.com/hndjmq



Two similarly sized planets collide.



The bodies are sheared by the impact.



Within hours, the debris heats up, mixes and begins to collapse.

► the inner Solar System do, then why is the composition of the Moon so similar to the outer portions of our planet?

So far, the solutions proposed appeal to extra processes — such as extensive mixing of materials from both bodies or a later gravitational resonance with the Sun — the feasibility of which are unclear. Planetary scientists need to double modelling efforts of the Earth–Moon system and identify chemical signatures in lunar and terrestrial rocks that could rule out some scenarios or suggest alternatives.

MORE ALIKE THAN DIFFERENT

There are clear differences in the compositions of the Moon and Earth. Earth's core is rich in iron, which comprises about 30% of the planet's mass. By contrast, iron contributes less than 10% to the mass of the Moon. The Moon is also less rich in elements that vaporize readily, such as potassium, suggesting that they may have boiled off and been lost as the Moon formed from the hot disk.

Analyses of samples brought back by the Apollo missions in the 1970s have shown that the silicate mantles of the Moon and Earth share identical oxygen isotope compositions (to within measurement precision)¹, distinct from those of meteorites from Mars and from most of the asteroid belt. In recent years the similarities have mounted. The chromium, titanium, tungsten and silicon isotope compositions of the Moon and Earth now also seem to be indistinguishable^{2–4}.

Gravity observations of the Moon from NASA's Gravity Recovery and Interior Laboratory (GRAIL) spacecraft, combined with topography data from NASA's Lunar Reconnaissance Orbiter, have reduced estimates for the thickness of the Moon's crust and its aluminium abundance. These measurements suggest that refractory elements (metals with high condensation temperatures) are similarly abundant in both bodies⁵, rather than more prevalent in the Moon, as previously thought.

Collectively, these data imply that either the Moon formed from material originating directly from Earth's mantle, or that the Moon

and the silicate portion of Earth each formed from an identical mix of material. Special circumstances seem to be required in either case.

IMPACT MODELS

Moon-forming collisions are studied through simulations. Because the energy caused by the impact of the colliding planets is high enough to melt or even partially vaporize them, pressure forces and phase changes are incorporated into the models. Gravitational interactions and torques are also included because the collision distorts the planets and ejects debris into a disk. Mantle and core materials need to be tracked.

In the canonical giant-impact model, developed since the late 1970s, the Moon is explained as the product of a slow, glancing blow from a Mars-sized body — about 10–15% of Earth's mass — on the early Earth⁶. The collision left Earth spinning rapidly, once every five hours, with the Moon orbiting close to Earth. Gravitational interactions and torques then caused the Moon's orbit to expand and Earth's rotation to slow to our current 24-hour day. This model is consistent with the Moon's mass, its lack of iron and the angular

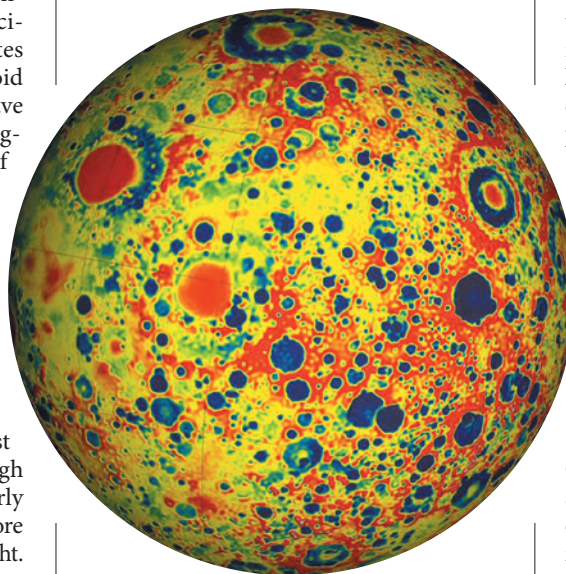
momentum of the Earth–Moon system.

More detailed chemical properties are harder to explain. The giant-impact model has the Moon condensing from material in a disk, which in canonical impacts is derived primarily from the impactor's mantle. But it is improbable that the impactor had the same composition as the early Earth. The oxygen isotope composition of Mars, for example, differs from that of the Earth by more than a factor of 50 (ref. 1). If the impactor was as different from Earth as Mars is, its signature would still be detectable in the Moon, even after a giant collision.

An elegant solution, known as equilibration, was proposed in 2007 by planetary scientists Kaveh Pahlevan and David Stevenson⁷. They suggested that vapour from the disk and the outer Earth mixed after the impact but before the Moon formed. But there are difficulties with this proposal. It takes at least 100 years for vapour from the disk and Earth to diffuse and mix thoroughly. But in that time the distant portions of the disk should have begun to coalesce into the Moon⁸.

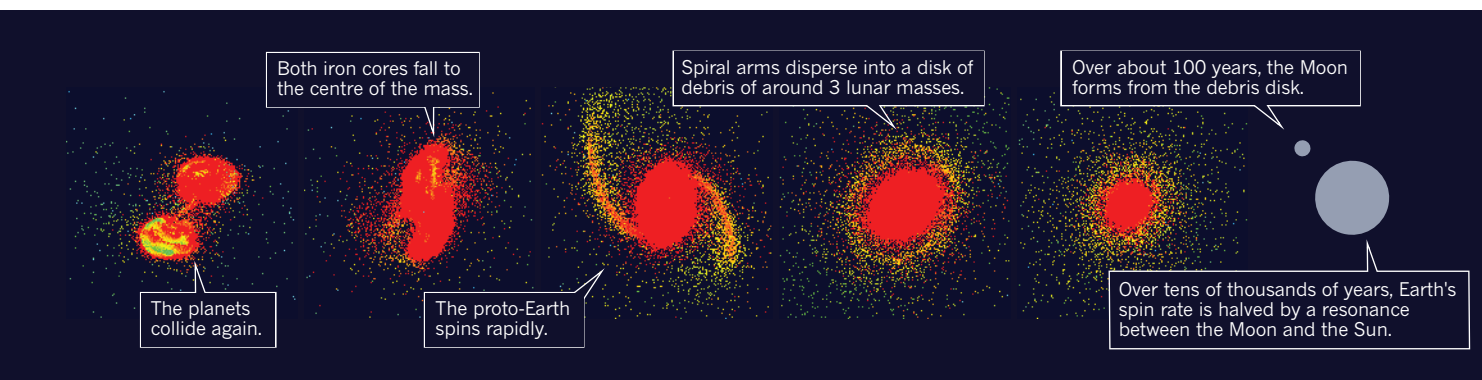
It is possible that the inner portions of the Moon could have retained the composition of the impactor. There are no signatures of this in lunar rocks; however, these represent only the outermost few hundred kilometres of the satellite's interior. Another conundrum is that volatile elements in the post-collision vapour would be expected to mix more readily than refractory ones, yet both oxygen and titanium, for instance, are identical in the two bodies.

In 2012, planetary scientists Matija Ćuk and Sarah Stewart⁹ broadened the range of possible Moon-forming impacts. Earth's oblate shape causes the orientation of the Moon's elliptical orbit to gradually rotate with a period that lengthens as the Moon's orbit expands — a process known as precession. Ćuk and Stewart showed that a resonant state between the Moon and the Sun that occurs when the lunar precession period matches the one-year period of Earth's orbit could — if it persists long enough — halve Earth's spin rate. Impacts of higher angular momentum then become viable, including



The Moon's gravity field as mapped by NASA's Gravity Recovery and Interior Laboratory.

NASA/JPL-CALTECH/MIT/GSFC



two cases that can produce a disk with the composition of Earth's mantle.

The 'fast-spinning Earth' scenario, proposed by Čuk and Stewart⁹, invokes the collision of an object slightly smaller than Mars with an Earth that is already rotating with a 2–2.5-hour day owing to a previous large impact. Because Earth is spinning close to the critical rate at which it becomes unstable, the Moon-forming impact ejects part of Earth's mantle into orbit, leading to a disk.

Also in 2012, I proposed the 'half-Earth impactor' scenario¹⁰. Here, the Moon arises from a collision between two planets, each of about half of Earth's mass (see 'How to make a Moon'). Both final planet and disk then comprise about half impactor and half target material. This model is simpler than the fast-spinning-Earth model because it does not require a specific prior large impact. But it demands a large impactor, and so may still be less probable than the canonical impact.

Both 2012 models account for the similar oxygen, chromium and titanium compositions of the Moon and Earth. To explain similarities in silicon and tungsten — elements that interact with metals — both models require that the impactor's iron core remains largely intact as it descends through Earth's mantle to merge with Earth's core, avoiding substantial metal–silicate interactions. But it remains unclear whether the resonance mechanism needed to slow Earth's rotation in these more extreme scenarios is likely or requires an improbably narrow range of conditions. In other words, is the origin of our Moon a rarer event than we believed, or are we missing something?

FUTURE DIRECTIONS

Lunar-origin studies are in flux. No current impact model stands out as more compelling than the rest. Progress in several areas is needed to rule out some theories, support others or direct us to new ones.

First, a better understanding of what happened between the formation of the disk and the accumulation of the Moon from the disk is essential, because this phase established the Moon's properties. Did mixing homogenize the composition of the disk and

the planet before the Moon formed? Were volatile elements lost from the disk, and, if so, did the pattern of loss vary with the disk's temperature? Canonical impacts produce a mostly liquid disk whereas in the high-angular-momentum impacts, the disks are initially largely vapour. Such disk-evolution models are technically challenging and will require a multidisciplinary approach incorporating both dynamics and chemistry.

Second, the likelihood that a resonance altered the Earth–Moon angular momentum needs to be assessed for a variety of physical states of the early Earth and Moon and using state-of-the-art models for the tidal interactions between them.

Finally, further isotopic comparisons of lunar and terrestrial materials would be extremely valuable. They should include highly refractory elements, such as calcium, to test the equilibration model. Finding that an element that could not have mixed in a vapour phase in 100 years is the same in the Moon and Earth but different in Mars would argue against equilibration; finding Earth–Moon isotopic differences in such a highly refractory element would support it.

Oxygen provides arguably the most important isotopic constraint on lunar formation. The distinct oxygen isotopic compositions of the Earth–Moon system, Mars and most meteorites reflect different initial compositional reservoirs in the inner Solar System. This simplifies the interpretation of oxygen compositions compared with elements such as silicon, whose isotopic abundances are affected by later planet-forming processes (such as crustal extraction). Increasing the precision of oxygen isotope measurements could potentially rule out some impact scenarios.

It remains troubling that all of the current impact models invoke a process after the impact to effectively erase a primary outcome of the event — either by changing the disk's composition through mixing for the canonical impact, or by changing Earth's spin rate for the high-angular-momentum narratives.

Sequences of events do occur in nature, and yet we strive to avoid such complexity in our models. We seek the simplest possible

solution, as a matter of scientific aesthetics and because simple solutions are often more probable. As the number of steps increases, the likelihood of a particular sequence decreases. Current impact models are more complex and seem less probable than the original giant-impact concept.

A clue may lie in Venus. The assumption that the Moon-forming impactor had a composition very different from that of Earth is

"Is the origin of our Moon a rarer event than we believed, or are we missing something?"

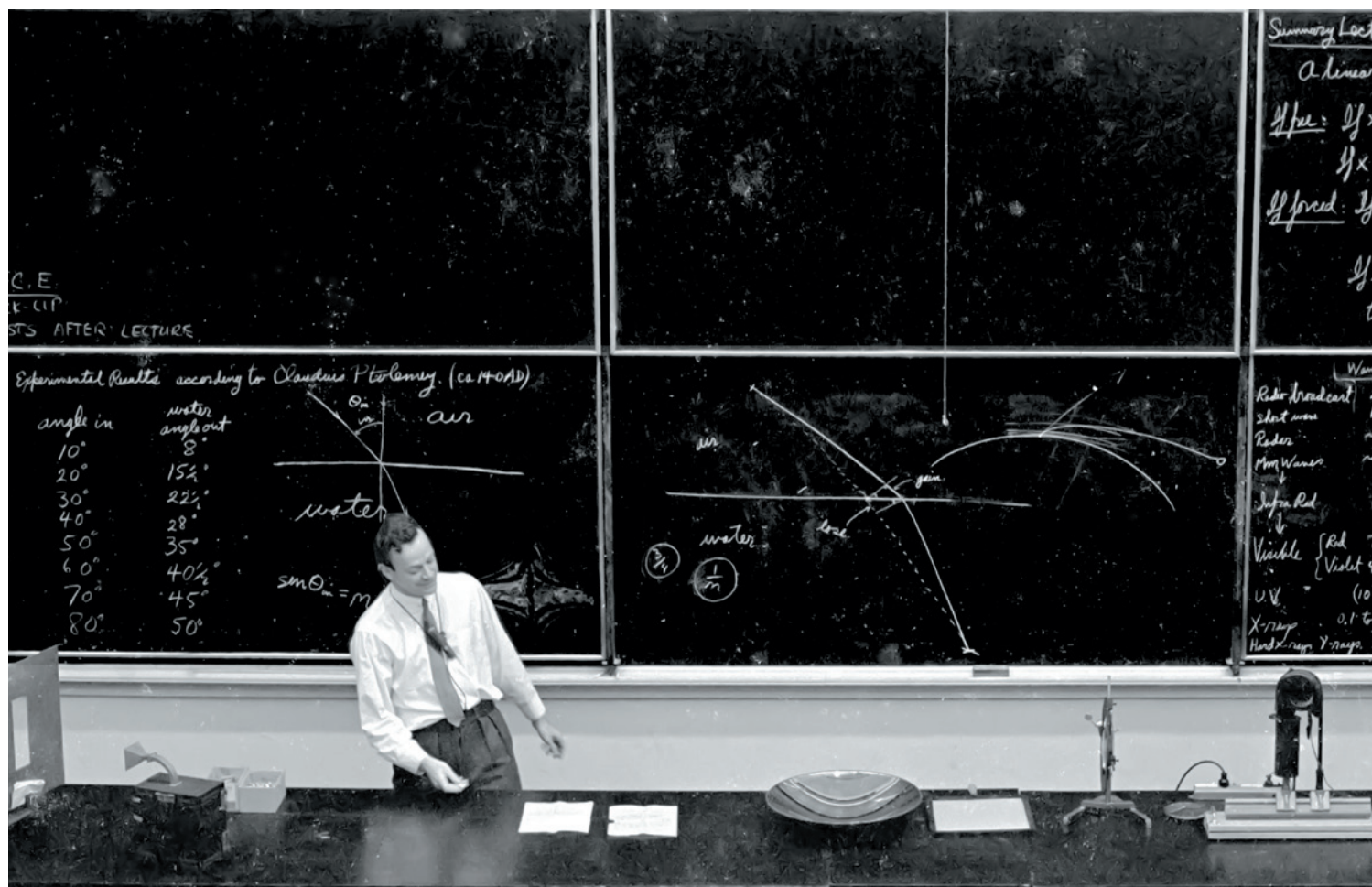
largely based on what we know about Mars.

We do not know the isotopic composition of Venus, the planet most similar to Earth in both mass and distance from the Sun. If Venus's composition proves similar to that of Earth and the Moon, Mars would then seem to be an outlier, and an impactor composition akin to Earth's would be more probable, removing many objections to the canonical impact.

Determining the isotopic composition of Venus's key elements will probably require a mission to the planet. Such a tantalizing prospect reminds us how much there is still to learn in our Solar System backyard. ■ **SEE NEWS & VIEWS P.90**

Robin Canup is associate vice-president of the Planetary Science Directorate of Southwest Research Institute, Boulder, Colorado.
e-mail: robin@boulder.swri.edu

1. Wiechert, U. *et al. Science* **294**, 345–348 (2001).
2. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. *Nature* **450**, 1206–1209 (2007).
3. Armytage, R. M. G., Georg, R. B., Williams, H. M. & Halliday, A. N. *Geochim. Cosmochim. Acta* **77**, 504–514 (2012).
4. Zhang, J., Dauphas, N., Davis, A. M., Leya, I. & Fedkin, A. *Nature Geosci.* **5**, 251–255 (2012).
5. Wiczorek, M. A. *et al. Science* **339**, 671–675 (2013).
6. Canup, R. M. & Asphaug, E. *Nature* **412**, 708–712 (2001).
7. Pahlevan, K. & Stevenson, D. J. *Earth Planet. Sci. Lett.* **262**, 438–449 (2007).
8. Salmon, J. & Canup, R. M. *Astrophys. J.* **760**, 83 (2012).
9. Čuk, M. & Stewart, S. T. *Science* **338**, 1047–1052 (2012).
10. Canup, R. M. *Science* **338**, 1052–1055 (2012).



Richard Feynman lecturing in 1962 on optics and Pierre de Fermat's principle of least time.

IN RETROSPECT

The Feynman Lectures on Physics

Rob Phillips celebrates the US physicist's seminal series as it nears its 50th anniversary.

Over the past three decades, I have asked hundreds of people to name the five or ten books that have meant the most to them. Although Jane Austen's *Pride and Prejudice* tops the list, *The Feynman Lectures on Physics* is the science title most often cited. That may say something about the kind of readers I talk to, but it is an accurate reflection of the broad reach of this half-century-old scientific classic.

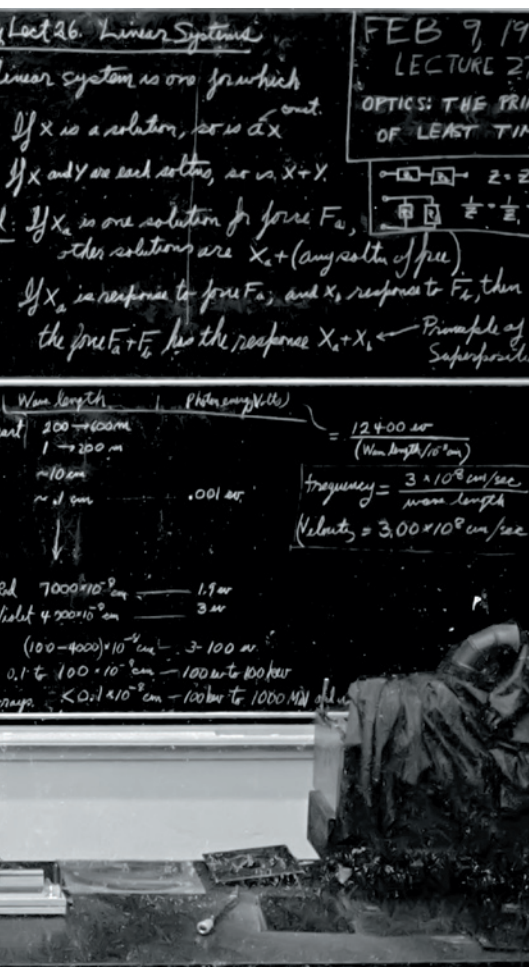
The book was based on a course the Nobel-prizewinning theoretical physicist and polymath Richard Feynman taught from 1961 to 1963, in an attempt to reinvigorate 'freshman physics' at the California Institute of Technology (Caltech) in Pasadena. In 1964, the course was published as the three-volume *The Feynman Lectures on Physics*, by Feynman and fellow physicists Matthew Sands

and Robert Leighton. With his lectures, Feynman joined a long tradition of famed physicists — such as Max Planck, Arnold Sommerfeld, Wolfgang Pauli and Lev Landau — providing personal grand vistas. Unlike those, Feynman's vista is 'elementary' and joyous — a joy deeply magnified in the audio version.

What makes these lectures timeless? Elementary physics has been taught to undergraduates for nearly a century with relatively little change. Over the past 50 years the subject has been even more static. Textbooks and introductory courses have largely targeted those planning to study medicine and engineers with a focus on formulaic problem-solving and exam preparation, rather than cultivating a wonder for nature and the development of physical intuition.

Superficially, Feynman's primer touches on the same topics that others do: mechanics, thermodynamics, optics, electricity and magnetism, and modern physics. Beneath this veneer of common cause, his introduction to elementary physics seems to have higher aspirations — the love of nature and a grasp of it through experimentation and reasoning. In Feynman's hands, even a topic as mundane as projectile motion becomes the story of how Galileo and Newton unlocked the secrets of planetary motion. Feynman's physics is about simplicity, beauty, unity and analogy, presented with enthusiasm and insight that bursts from the page.

He works this magic even in areas often thought to be the most boring parts of the curriculum. For example, his fascination with the way that Newton's second law of



TOM HARVEY/CALIFORNIA INSTITUTE OF TECHNOLOGY

motion, $F = ma$, can describe the motions of large, composite objects such as galaxies leads intuitively to the profound idea of the centre of mass. Feynman also repeatedly appeals to 'variational' principles based on minimizing quantities such as travel time (pictured). This is seen nowhere more impressively than in the way he develops optics by thinking about the transit of light rays as they pass through various media, whether lenses or the atmosphere. These same ideas return in his treatment of the elliptical motions of planets. When talking about Brownian motion (the random movement of particles in a gas or liquid as they collide with molecules of that medium), he elegantly teaches us the fluctuation-dissipation theorem, which relates how rapidly particles diffuse to the drag force they experience, without ever naming it as such. And he similarly provides an advanced but accessible introduction to elasticity — the likes of which, unfortunately, advanced physics students rarely see even now.

NATURE.COM
Daniel Cressey
on the art of the
Feynman diagram:
go.nature.com/d8eikf

In a handful of pages on electrostatic analogs and the unity of nature, Feynman points out how many different phenomena

can all be approached using the same underlying mathematics — the electrical potential around charged objects, heat flow between plates held at different temperatures, the vibrations of a drumhead, the diffusion of neutrons and the flow of a fluid past a sphere. Such unity is further revealed by his potpourri of examples of resonance in nature. These range from the 'oscillator' of Earth's atmosphere as it sloshes back and forth, driven by the Moon, to the Mössbauer effect revealed when atomic nuclei absorb high-frequency radiation.

Feynman seems to be teaching the idea that there is no one right way to view a problem. As he put it in his 1965 Nobel lecture, "Theories of the known, which are described by different physical ideas may be equivalent in all their predictions and are hence scientifically indistinguishable. However, they are not psychologically identical when trying to move from that base into the unknown ... I, therefore, think that a good theoretical physicist today might find it useful to have a wide range of physical viewpoints and mathematical expressions of the same theory."

One of the most delightful features of the *Lectures* is that Feynman is constantly on the lookout for physics writ large. From lightning to the periodic table and the energy levels of chlorophyll, he is not interested in naming conventions that separate different academic disciplines. How many physics books have a section entitled "More organic chemistry"? In it, he shows us how to use simple quantum-mechanical models to work out the spectrum of energy levels of different types of molecules. Chapters 35 and 36 of Volume 1 take on topics related to vision, such as the anatomy of rod cells, how the molecule retinal in photoreceptor cells works, the resolution of the compound eye of the bee and the mysteries of colour vision — reflecting his 1960s adventures in biology. Feynman would relate these in more detail in *Surely You're Joking, Mr. Feynman! Adventures of a Curious Character* (W.W. Norton, 1985), his book of reminiscences based on taped conversations with Ralph Leighton, the son of his *Lectures* co-author Robert.

BREAKING THE MOULD

The breadth of Feynman's scientific interests was brought home to me during a chance visit to the Caltech archives. Accompanying a colleague who wanted to examine the papers of Max Delbrück, I noticed two boxes of papers open on a desk. They turned out to be Feynman's. Randomly flicking through, I was struck by nearly 100 pages of notes from the 1960s with a peculiar, decade-long timeline marked in Feynman's unmistakable writing. For March 1966, he had written "Footprints of Tumor Viruses, The Nerve Axon." For April 1966, "Sex differences in the brain, Chromosome analysis by computer, antibiotics & the

genetic code". As I guessed and later confirmed, these were all topics he was reading about in *Scientific American*.

Feynman seems to have been hard at work learning anything and everything he could about biology, coloured by physical reasoning — although by then he was one of the most famous physicists in the world. Page after page is littered with Feynman's drawings, notes and questions on topics ranging from the phylogeny of plants and animals, and the structure of proteins to the beautiful membrane structures of mitochondria. They give a feeling of his roving mind busily formulat-

"Feynman's physics is about simplicity, beauty, unity and analogy, presented with enthusiasm and insight."

ing his own version of biology. More importantly, they show once again his delight in learning about the marvels of nature and his urge to bring order to the things he knew. At the end

of these notes, Feynman returns to the mysteries of the quantum world — his biological musings are replaced by lengthy calculations.

Mark Twain quipped that a classic is something that everybody wants to have read and nobody wants to read. Feynman's classic breaks the mould. Some respondents to my 'favourite books' query speak of dog-eared copies lovingly read on long stints around the globe. Among them are a young high-school student who in Yugoslavia's Communist era tried to master Maxwell's equations; an Israeli army officer stealing time to read every page over years of duty; and a brilliant Indian undergraduate trying to breathe life into a freshman physics course designed to 'train' engineers. One travel-bum mathematics student decided it was time to learn physics — and turned to the lectures, eventually landing a place as a graduate student in Feynman's former department at Caltech. The book has a cult following among non-specialist readers as well.

As Feynman writes in his epilogue to the series: "I most wanted to give you some appreciation of the wonderful world and the physicist's way of looking at it ... it is even possible that you may want to join in the greatest adventure that the human mind has ever begun." It is because they serve as an expert and loving guide to this great adventure that *The Feynman Lectures on Physics* are as timely now as they have ever been. ■

Rob Phillips is the Fred and Nancy Morris Professor of Biophysics and Biology at the California Institute of Technology.
e-mail: phillips@pboc.caltech.edu

The New Millennium Edition of the *Lectures* is available from Basic Books. The series is free at www.feynmanlectures.caltech.edu.



Streaming children into subjects on the basis of genetic testing may not be universally popular.

EDUCATION

Genetics in the schoolroom

Erika Check Hayden ponders a call for schools to embrace genetic information as a priority.

How should the ideal school be designed? A provocative book proposes that we start with genetics. In *G is For Genes*, Robert Plomin, the geneticist who heads the long-running Twins Early Development Study at King's College London, argues that his research is revealing the unfeasibility of all children thriving under 'one-size-fits-all' education regimes. With educational researcher Kathryn Asbury, Plomin lays out the case for a genetically influenced school scheme.

The data from the twin studies, Plomin and Asbury reveal, support the idea that our genes predispose us to excel or to lag behind in particular areas. Education can level the societal playing field by standardizing the learning environment, giving children the chance to

fulfil their genetic potential, but will never enable all children to perform equally well in all subjects. Instead, the authors argue, educators should help each child to reach a basic standard of performance in core areas such as reading and maths. Each child's education could then be personalized so that he or she can develop the skill set favoured by their genes — be that in academic subjects, sport, music or horticulture.

Plomin and Asbury acknowledge that personalizing schools would cost money, but say that teachers could already tailor lessons to a range of skill sets by using computer-based instruction programmed to adapt to students' performance. A school that strives to be sensitive to every child's individuality is the holy grail of most parents. But parents

— and scientists — might look more sceptically on the book's proposal to genotype all children with a "Learning Chip", which would assess the status of genetic markers relevant to learning potential, to gauge how a child's genes might influence their abilities.

At present, science seems some way off understanding enough about how genes interact with the environment and with each other to predict their influence on complex personality traits, such as an inclination towards maths. Such traits seem to involve many genes, each of which has a small influence, making them difficult to find. The largest-ever study of academic careers, for instance, enlisted 125,000 people, but found only three genetic traits that affect an individual's duration of schooling (a proxy for student achievement) by a few months. And even if such traits are found, predicting how they might influence a student's academic career will be difficult (see C. A. Rietveld *et al. Science* **340**, 1467–1471; 2013).

To be fair, Plomin and Asbury realize that the use of Learning Chips is not possible at the moment, but they seem confident of its future feasibility. However, the central idea of streaming children into certain subject areas on the basis of genetic tests may not find much support.

The authors argue that such tests could be used to identify students who need help to meet basic standards, and to help in the formulation of interventions. Their views are influencing policy: Dominic Cummings, an adviser to British education secretary Michael Gove, once invited Plomin to brief the UK Department for Education on the science of genetics. Cummings caused a firestorm in October with the release of a paper he authored that seemed to favour research into genetics in education.

The study of achievement and intelligence has attracted controversy because it is so often aimed at understanding the make-up and needs of the super-bright, such as students who perform well on standardized tests. They already have a societal advantage. So if it ever becomes feasible to use genes to predict achievement, geneticists such as Plomin face a big task. They will need to present a compelling case that they intend to use genetics for the good of all, not just for the continued advancement of the genetically favoured. ■

Erika Check Hayden is a writer for *Nature* in San Francisco, California.



G is For Genes:
The Impact of
Genetics on
Education and
Achievement

KATHRYN ASBURY
AND ROBERT PLOMIN
Wiley-Blackwell: 2013.

BLEND IMAGES/ALAMY

Correspondence

Take more care over glacier facts

We find it most unfortunate that as a leading journal you introduced crucial factual errors into an article on Himalayan glaciers (see J. R. Laghari *Nature* **502**, 617–618; 2013, and Correction *Nature* **503**, 464; 2013), particularly in light of lessons learned from the ‘glaciergate’ scandal that threatened the reputation of the Intergovernmental Panel on Climate Change (IPCC) in 2010.

Glaciers seem to be peculiarly at risk from careless errors (J. G. Cogley *et al.* *Science* **327**, 522; 2010, and J. S. Kargel *et al.* *The Cryosphere* **6**, 533–537; 2012), as well as from climate change. Editors should follow the IPCC’s example and sharpen up their fact-checking procedures. **Alex S. Gardner** *Clark University, Worcester, Massachusetts, USA.* agardner@clarku.edu

**On behalf of 4 co-signatories (see go.nature.com/gx2t9y for full list).*

Recycle waste for nourishing soils

On World Soil Day (5 December), it is worth pointing out that there are opportunities for sustainable soil management beyond the farm gates (*Nature* **502**, 607; 2013). We need to rethink modern society’s complex product flows so that waste and recycling practices can be adapted for enriching soil, which is a non-renewable resource.

Organic agriculture in subsistence-farming systems throughout the developing world requires inputs that are in short supply because of limited land, nutrients and organic matter. This deficiency could be remedied by importing nutrient sources and organic matter from other industrial sectors within national boundaries.

Processed industrial and domestic waste could contribute large amounts of nitrogen and phosphorus derivatives to soil,

for example, and human urine or slaughterhouse residues could be used to produce indigenous and highly efficient biofertilizers.

Johannes Lehmann *Cornell University, Ithaca, New York, USA.* cl273@cornell.edu

Please pass the microbes

Living and working among Tanzania’s Hadzabe people — one of the world’s last remaining hunter-gatherer groups — I witnessed the extraordinarily intimate relationship they share with microbes in their environment. This potentially provides them with a health-enriching source of gut microbial diversity, lost long ago in the modern lifestyle of the developed world.

Like the Hadzabe, all humans were presumably once connected to a huge microbial metacommunity through the guts, skin and feathers of animals in their territory. As well as sharing water sources tainted with the urine and faeces of animals as diverse as zebras, giraffes and bush pigs, the Hadzabe often consume the uncooked stomachs and colons of killed animals. They also ‘clean’ their hands in the animals’ partially digested and microbe-laden stomach contents (pictured), helping to transfer microbes among community members.

The lower diversity of gut microbes among populations

in the developed world (see, for example, T. Yatsunenko *et al.* *Nature* **486**, 222–227; 2012) may increase our susceptibility to opportunistic pathogens and diseases. We should be exploring the value of the Hadzabe people’s rich microbial sources, notwithstanding fundamental issues of sanitation and hygiene. **Jeff Leach** *Human Food Project, New Orleans, USA, and London School of Hygiene & Tropical Medicine, UK.*

jeff@humanfoodproject.com

All journals need to correct errors

Innocent errors are more commonplace than fraud in research papers, so they need to be identified and corrected promptly. We call for more journals to accept their responsibility to ensure that this happens.

We searched for errors in 107 papers in the fields of engineering, materials and computer science, which were based on existing small data sets (see G. Taguchi *et al.* *Quality Engineering Handbook*, Wiley; 2004). Our search revealed an alarming number of errors. Ten papers had one or more mistakes that were substantial enough to affect the findings and conclusions. There were errors that were not so significant in almost one-third of the papers.

We notified the journals that published substantial errors. Only four of the ten formally corrected their mistakes. These corrections were published 4–8 months from initial notification. The remaining journals declined to publish a correction; some even had a policy not to publish criticisms of their papers.

Journals that were willing to print general corrections (through errata, letters and notes) were found to have corrected 0.11–0.71% of their existing papers. This represents less than 10% of the error rate

exposed here (see also R. D. Chirico *et al.* *J. Chem. Eng. Data* **58**, 2699–2716; 2013).

Although our selection of journals is narrow, our findings hint that many crucial errors may go forever uncorrected. The risk is compounded by the rarity of attempted replication studies (see go.nature.com/dstij3).

It should be standard practice for journals to insist on full data provision by authors and on declarations of individual author contributions (see, for example, go.nature.com/lwkkqo). Prolonged investigation may be necessary in cases of suspected fraud — but honest errors can be corrected relatively quickly. And this demands a consistent correction policy among journals.

Jonathan D. Linton *University of Ottawa, Canada.* linton@uottawa.ca

**On behalf of 5 co-signatories (see go.nature.com/ew2i9i for full list).*

Doctor Who and the ageing enigma

In his time-travel musings to mark *Doctor Who*’s half-centenary, Andrew Jaffe overlooks one facet of the Gallifreyan narrative: regeneration (*Nature* **502**, 620–622; 2013). Rather than die, time lords can choose to regenerate (purportedly up to 12 times) once their bodies are worn and weary.

Even though time travel is still in the realm of fantasy, progress in the sphere of rejuvenation is palpable. Many of the disagreeable features of ageing are now successfully countered by medical science and technology.

The Doctor would surely advise us to focus more closely on preventing degeneration, especially in light of the boom in life expectancy over the past 50 years and its accompanying social and scientific challenges.

Faisal R. Ali *University of Manchester, UK.* f.r.ali.01@cantab.net



Microbial matter as hand cleanser.

Leonard Herzenberg

(1931–2013)

Immunologist who pioneered cell-sorting technology.

Leonard Herzenberg, together with his wife and scientific partner of more than 60 years, Leonore, transformed immunology. His accomplishments continue to influence every aspect of modern biological science.

Were it not for Herzenberg, researchers might still be waiting for fast, precise ways to count and sort cells. That task, now routine, is essential for working out how cancers and tissues grow, and for diagnosing certain diseases. Herzenberg also developed technologies to create monoclonal antibodies — lab-produced versions of the proteins used by the immune system to recognize dangerous cells and toxins. These are now deployed in laboratories for innumerable assays, and in humans for certain cancers, infections and inflammatory diseases.

Herzenberg, who died on 27 October aged 81, was born in New York City. After graduating in 1952 from Brooklyn College in New York, he began doctoral studies in genetics at the California Institute of Technology in Pasadena, joining a department that included seven future Nobel laureates. Among this auspicious group, Herzenberg blossomed scientifically and, along with Leonore, whom he had met at Brooklyn College, he also became politically active. Len and Lee, as they were known, worked with a chemist down the hall — Linus Pauling, who himself would go on to win two Nobel prizes — to start a local chapter of the Federation of American Scientists that worked to combat the insidious efforts of McCarthyism in the 1950s, among other goals.

After Len received his PhD, the Herzenebergs moved to Paris, where Len did a fellowship at the Pasteur Institute with biologist and Nobel laureate Jacques Monod. Discussions at lunch revolved alternately around bacterial genetics and the French Resistance. While abroad, Herzenberg was drafted into the US peacetime army. He arranged to serve in the Public Health Service and moved to Bethesda, Maryland, to join the lab of pathologist Harry Eagle at the US National Institutes of Health. Here, Herzenberg “carried a pipette rather than a gun” for his country. He helped to define the minimal components of the nutrient broth necessary to clone cells. More than a decade later, this enabled him to produce antibodies in cell culture, a technology that generates billions of dollars’ worth of drugs and lab reagents.



In the 1960s, after moving to California to work at Stanford University, Herzenberg recognized the need for an automated, high-throughput method to enumerate and separate rare cells in a population of millions. Such technology could be used, for example, to explore whether individual immune cells made one or several kinds of antibodies, or to identify those producing antibodies capable of killing cancer cells.

Herzenberg formed a team of engineers and biologists to take a machine that sorted particles by size and enhance it with techniques to detect fluorescence and manipulate droplets. The result was modern flow cytometry — a way to count and separate viable cells in a stream of fluid lit by lasers that reveal fluorescent markers on the cells. Fluorescence-activated cell sorting (known as FACS) is now widely used in labs and hospitals. For instance, FACS can isolate fetal cells from maternal blood, which allows clinicians to perform genetic testing using a minimally invasive procedure. Remarkably, the fundamental aspects of FACS have not changed in the more than 40 years since its invention.

Another ubiquitous technology also owes its existence to the Herzenebergs. In 1976, on a sabbatical at the University of Cambridge, UK, the Herzenebergs worked with

biochemist César Milstein on his technology to fuse cells. They created cell lines that produce made-to-order antibodies which attach to specific markers on cell surfaces.

Len recognized the power of this technology to identify specific proteins and cell populations. Lee coined the term ‘hybridoma’ for a class of workhorse lab cells, a combination of an antibody-producing white blood cell and an immortal tumour cell.

Although Herzenberg’s patents are among Stanford’s highest earning, he believed that any scientific advance belonged to the people who had funded it; that is, to the public, whose taxes supported his grants. He asked co-inventors of flow cytometry to assign patent royalties back to the lab. He freely shared resources that his laboratory developed, providing researchers from other labs with reagents, cell lines, data and information, sometimes well before publication. This collaborative philosophy was appreciated by scientists worldwide. It also helped to accomplish his primary goal: more-rapid advances.

Throughout his career, Herzenberg was actively involved in sociopolitical causes. His early exposure to the dangers of the anti-intellectualism and misinformation of McCarthyism seeded a lifelong commitment to bringing rational thought to public discourse. He railed against proponents of eugenics and nuclear proliferation, and, in later years, against those who stigmatized people with HIV.

He also supported efforts to promote career opportunities and equality for women and minorities in science. The Herzenebergs established programmes to bring disadvantaged young people in California’s San Francisco Bay Area to Stanford to learn about medical research. Lee continues this tradition of balancing outreach with scientific research.

Although Len is considered the father of modern flow cytometry, his contributions extend far beyond that. He was a great scientist, a great collaborator and a great mentor. ■

Mario Roederer is senior investigator in the ImmunoTechnology Section, Vaccine Research Center, US National Institute of Allergy and Infectious Diseases, Bethesda, Maryland. He was in the Herzenebergs’ laboratory from 1988 to 1999.
e-mail: roederer@nih.gov

natureINSIGHT

COASTAL REGIONS

5 December 2013 / Vol 504 / Issue No 7478



Cover image
Norman Kuring

Editor, Nature
Philip Campbell
Publishing
Richard Hughes
Production Editor
Jenny Rooke
Art Editor
Nik Spencer
Sponsorship
Reya Silao
Production
Ian Pope
Marketing
Elena Woodstock
Steven Hurst
Editorial Assistant
Abbie Williams

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com



nature publishing group

Earth's coastlines are a mere sliver between the continental interiors and the open ocean, yet their importance for humans is larger than their surface area suggests. Coastal areas house billions of people and provide our societies with multiple ecosystem services that have environmental and economic value. The intense human usage of coastal areas, for example through land-use change, groundwater extraction and dam construction, is placing pressures on coastal ecosystem functions that may be exacerbated by global increases in temperature and sea level. The articles in this Insight explore some of the local and global processes that are shaping coastal systems, and how they sustain diverse, but mostly human-dominated, ecosystems.

A variety of processes, operating at a range of spatial scales, govern the stability of coastal systems. For instance, ocean circulation, climate and glacial hydrology are driving the submarine melting of Greenland's outlet glaciers. Wetland stability and coastal biogeochemical cycles are affected by local sediment and nutrient supply. Human activities have led to changes in these processes at all spatial scales, threatening the survival of coastal ecosystems.

However, careful management could help to ensure the sustainability of these important regions and may reveal additional ecosystem services. Coastal defences could be engineered through ecosystem creation and restoration on a large scale, providing more environmentally sound coastal protection than conventional engineering. Exploitation of offshore, submarine groundwater reserves might complement the water supply to densely populated coastal regions.

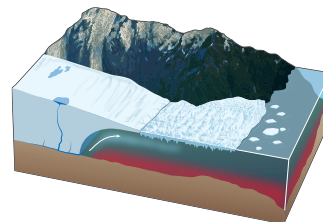
This Insight presents some of the many scientific challenges associated with coastal research. In so doing, it highlights the need for a better understanding of the vulnerability of the coastal landscape, which will be required to both protect the valuable coastal region and enjoy its services.

Juliane Mossinger, Michael White & Patrick Goymer
Senior Editors

CONTENTS

REVIEWS

- 36 North Atlantic warming and the retreat of Greenland's outlet glaciers**
Fiammetta Straneo & Patrick Heimbach



- 44 Coastal flooding by tropical cyclones and sea-level rise**
Jonathan D. Woodruff, Jennifer L. Irish & Suzana J. Camargo

- 53 Tidal wetland stability in the face of human impacts and sea-level rise**
Matthew L. Kirwan & J. Patrick Megonigal



- 61 The changing carbon cycle of the coastal ocean**
James E. Bauer, Wei-Jun Cai, Peter A. Raymond, Thomas S. Bianchi, Charles S. Hopkinson & Pierre A. G. Regnier

- 71 Offshore fresh groundwater reserves as a global phenomenon**
Vincent E. A. Post, Jacobus Groen, Henk Kooi, Mark Person, Shemin Ge & W. Mike Edmunds

PERSPECTIVES

- 79 Ecosystem-based coastal defence in the face of global change**
Stijn Temmerman, Patrick Meire, Tjeerd J. Bouma, Peter M. J. Herman, Tom Ysebaert & Huib J. De Vriend
- 84 Green and golden seaweed tides on the rise**
Victor Smetacek & Adriana Zingone

North Atlantic warming and the retreat of Greenland's outlet glaciers

Fiammetta Straneo¹ & Patrick Heimbach²

Mass loss from the Greenland ice sheet quadrupled over the past two decades, contributing a quarter of the observed global sea-level rise. Increased submarine melting is thought to have triggered the retreat of Greenland's outlet glaciers, which is partly responsible for the ice loss. However, the chain of events and physical processes remain elusive. Recent evidence suggests that an anomalous inflow of subtropical waters driven by atmospheric changes, multidecadal natural ocean variability and a long-term increase in the North Atlantic's upper ocean heat content since the 1950s all contributed to a warming of the subpolar North Atlantic. This led, in conjunction with increased runoff, to enhanced submarine glacier melting. Future climate projections raise the potential for continued increases in warming and ice-mass loss, with implications for sea level and climate.

Mass loss from the Greenland ice sheet (GrIS) quadrupled from 1992–2001 ($51 \pm 65 \text{ Gt yr}^{-1}$) to 2002–2011 ($211 \pm 37 \text{ Gt yr}^{-1}$), contributing to a rise in global mean sea level of $7.5 \pm 1.8 \text{ mm}$ from 1992 to 2011, roughly twice that from the Antarctic ice sheet^{1,2} (Box 1). At present, GrIS mass loss accounts for one quarter of the observed global sea-level rise³. Persistent ice loss from Greenland is also increasing the freshwater input into the North Atlantic. Conventionally, Greenland's freshwater discharge (Box 1) into the North Atlantic has been assumed to be negligible compared with the freshwater export from the Arctic Ocean⁴. However, a recent study⁵ argues that the cumulative freshwater anomaly discharged by the GrIS since 1995 amounts to a third of Arctic-origin freshwater anomalies that have disrupted dense water formation in the North Atlantic in the past. GrIS mass loss therefore may soon affect the Atlantic meridional overturning circulation (AMOC; Box 1), a key component of the climate system.

GrIS mass loss is due, in equal parts, to two processes^{6,7} (Fig. 1a). First, negative surface mass balance (Box 1) is attributed to a persistent increase in surface melt in southeast and west Greenland^{6,8}. Second, increased ice discharge resulted from the speed up, thinning and retreat of multiple marine-terminating glaciers (in contrast to those terminating on land⁹) in southeast and west Greenland that began in the mid-1990s^{10,11} (Fig. 2a) and spread to the northwest (Fig. 2b) in the mid-2000s¹². Whereas many of the southern glaciers have slowed down since their peak speeds in 2005, most continue to flow faster than in the mid-1990s, although there is variation between glaciers in the same area¹³.

The widespread and synchronous nature of changes in both surface mass balance and ice discharge are indicative of a response to external forcings, and consistent with observations of atmospheric¹⁴ and oceanic warming¹⁵ (Fig. 1b, c) over and around Greenland. The forcing responsible for the decrease in surface mass balance is evident^{6,8}: changes in precipitation and rising air temperatures have resulted in increased surface melting over the ice sheet^{14,16,17}. The mechanisms and forcings behind the increased ice discharge, however, remain elusive.

Glacier speed-up resulted from initial retreat^{11,18} of the marine termini (Fig. 2a, b) that decreased the resistance to ice flow, increased calving and thinning¹⁹, and led to further retreat^{20–24}. One leading hypothesis to explain the initial retreat of the glaciers involves an increase in submarine melting at the ice–ocean interface^{23,25,26}. Ocean forcing is also

invoked as a potential driver of changes in the ice mélange²⁷ (Fig. 2a). Finally, changes in sea-surface temperatures around Greenland have been correlated to changes in coastal air temperatures and, in turn, to changes in runoff⁶, implying that ocean-induced localized atmospheric changes may be affecting the GrIS.

When initially proposed, the above-mentioned hypotheses were based on the observation that the glaciers began to speed up and retreat at the same time as the waters off west Greenland warmed rapidly²⁸. The pervasive lack of ocean data near the glaciers at that time and our limited understanding of the mechanisms of ice-sheet–ocean interaction made it difficult to examine its plausibility. Since then, multiple field, modelling and theoretical studies have greatly advanced our understanding of how ocean variability affects the glaciers' margins.

In this Review we discuss these advances and conclude that warming of the subpolar North Atlantic (SPNA) ocean and atmosphere led to an increase in submarine melting of the glaciers. We discuss the context of warming of the SPNA (which is unprecedented in the historical record, except for a similar warm period in the 1930s) in terms of climate variability over the North Atlantic sector. We argue that the SPNA warming cannot be attributed to a single mode of atmospheric or oceanic variability. Rather, present-day warming is due to the superposition of these modes on a long-term ocean warming trend. This Review complements other articles on the ice sheet's response to ocean forcing^{23,25,26}, by focusing on the ocean dynamics around Greenland and its larger scale context.

From basin to boundary layer scales

The subject of ice-sheet–ocean interactions in Greenland described in this Review involves a range of scales, from the basin-wide North Atlantic (1,000 km scale) to the turbulent boundary layer at the ice–ocean interface (millimetre scale). We present the observational evidence for the physical processes that act on, and connect, these scales.

Continental shelf hydrographic variability

The increase in ice discharge that started in the mid-1990s is associated with the retreat of glaciers at the margins of the SPNA and its extension into Baffin Bay (between west Greenland and the Canadian Arctic Archipelago; Fig. 3a). The circulation around the SPNA is a cyclonic (anticlockwise) gyre with warm waters from the subtropics flowing

¹Department of Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. ²Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

BOX 1

Ice glossary

- **Units** Typical units used by the glaciological and oceanographic communities are mass loss in gigatonnes per year (Gt yr^{-1}), global mean eustatic sea-level rise in millimetres per year, and volume transport in sverdrups ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$). Conversion factors between these are: $1 \text{ Gt yr}^{-1} = 2.8 \times 10^{-3} \text{ mm yr}^{-1} = 3.17 \times 10^{-5} \text{ Sv}$.
- **Estuarine-driven circulation** The buoyancy-driven circulation associated with the entrainment of ambient water into the plume as it rises along the ice face and flows out of the fjord^{47,49}.
- **North Atlantic oscillation (NAO)** A leading pattern of atmospheric variability measuring the sea-level pressure differences between weather stations in the Azores and Iceland⁶⁶.
- **Atlantic multidecadal oscillation (AMO)** A mode of oceanic variability expressed as the sea-surface temperature (SST) anomaly over the North Atlantic^{70,71}.
- **Atlantic Meridional overturning circulation (AMOC)** Zonally

(west–east) integrated and vertically accumulated volume (or mass) transport in the Atlantic.

- **Freshwater discharge** The sum of solid ice discharge due to calving and liquid subglacial discharge from surface- and sub-glacial melting^{5,6}.

- **Ice mélange** A mixture of icebergs and sea ice found in front of the terminus of many Greenland glaciers that may affect calving²⁷.

- **Glacial isostatic adjustment (GIA)** Geodynamic and geodetic effects associated with ice-sheet mass loss, including visco-elastic rebound due to unloading of the mantle, adjustment of Earth's angular momentum and rotation, and change of Earth's gravity field (geoid) owing to mass redistribution⁸⁶.

- **Mass balances**

Total mass balance (MB) = surface MB (SMB) – discharge (D);
SMB = accumulation (A) – runoff (R); freshwater flux = R + D.

around the continental slopes of Greenland and North America encircling the colder, denser interior of the SPNA^{29,30}. Cold, fresh water from the Arctic flows around Greenland's 200–300 m deep continental shelves, partially buffering Greenland's coast from the warm, Atlantic waters³¹ (Fig. 3a). The glacier retreat coincided with a rapid warming of the SPNA that began in the mid-1990s^{15,29} and that continues today³² (Fig. 3b–d). The SPNA change is manifested in a warming of the upper 500–1,000 m of the waters off west Greenland, including the continental shelf^{28,33,34} (Fig. 1b) and extending to Baffin Bay³⁵. Data from the continental shelves of southeast Greenland are limited, but repeated annual hydrography across the Irminger Sea shows an extensive thickening of the Atlantic layer around the mid-1990s and suggests a similar warming of the shelf waters³⁰. The continental shelf warming is probably associated with more frequent intrusions of (warmer) Atlantic water in the deep troughs that stretch across Greenland's shelves³¹. Whether these have a surface signature³⁶ remains unclear.

Exchanges between fjord and continental shelf

Greenland's large marine-terminating glaciers are typically grounded several hundreds of meters below sea level at the head of long (10–100 km), narrow (<10 km) fjords that connect them to the continental shelf (which we refer to in this Review as shelf; Fig 4). Data from the fjords preceding the SPNA warming are too scarce to provide information that, in conjunction with recent surveys, could be used to describe how the fjords responded to the shelf warming. A comparison of ocean properties from two summer surveys taken before and after the mid-1990s warming (1993 and 2004) in Kangerdlugssuaq Fjord, southeast Greenland, shows warming of fjord waters³⁷, but it is unclear to what extent differences between the two short surveys are representative of longer term changes in the fjord compared with the large weekly to inter-annual variability. Recently collected data from several fjords, combined with dynamical considerations, however, provide a consistent picture that offers insight into how the fjord properties may have changed in response to the SPNA warming. Surveys have shown that the fjords contain a thick layer of warm (0–4 °C, compared with the freezing point of seawater, roughly –1.9 °C) saline, subsurface Atlantic water beneath a layer of cold, fresh polar water (Fig. 4)^{28,37–40}. The warmest Atlantic water is found in glacial fjords abutting the SPNA and Baffin Bay, whereas the coldest is found in glacial fjords abutting the Arctic Ocean in northern Greenland. This is consistent with variations in the mean Atlantic water properties on the nearby shelf and slope and reflects the distance (along its mean flow pathway) from the subtropical source region⁴¹. Along-fjord variations are relatively small, suggesting that Atlantic water is also found in the vicinity of the glaciers, although

most of the surveys terminate about 10 km from the glaciers' edge because of the inaccessibility of this region (Fig. 2).

The similarities between the fjord and the shelf properties are consistent with the fact that the fjords typically have deep (>200 m) sills that allow for a relatively unobstructed exchange between the two^{36,40,41}. This suggests that these fjords contained Atlantic water and

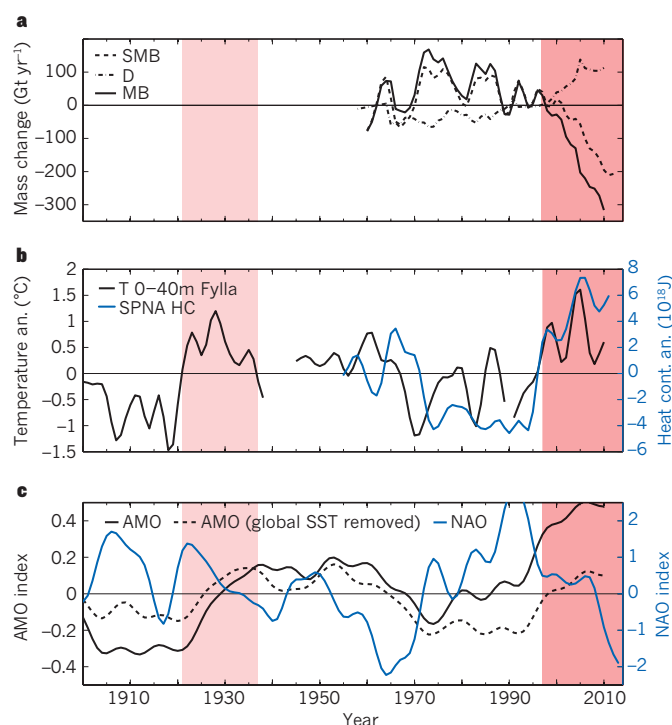


Figure 1 | Retreat of Greenland's outlet glaciers is occurring at a time when the waters of the subpolar North Atlantic are the warmest on record. **a**, Mass balance (MB), surface mass balance (SMB) and ice discharge (D) anomalies in gigatonnes per year based on refs 5, 6. **b**, Mean temperature anomaly (an.) of the upper 40 m at Fylla Bank, west Greenland⁵⁸ and heat content anomaly of the SPNA's upper 700 m⁵⁵. **c**, Atlantic multidecadal oscillation (AMO) index anomalies with and without the global SST trend^{71,73}, and North Atlantic oscillation (NAO) winter index⁶⁶. All time series have been extended to 2010 and 5-year low-pass filtered, and the mean with respect to the period shown has been removed. Recent glacier acceleration began in the late 1990s (dark shading), a similarly warm period occurred in the 1930s (light shading) with some evidence for glacier retreat of comparable magnitude.

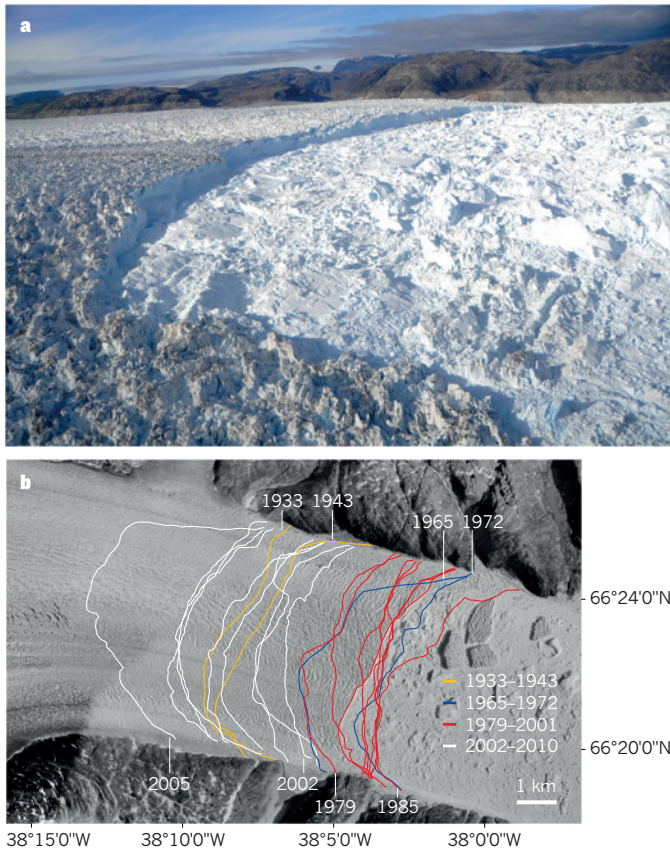


Figure 2 | Retreat and thinning of Greenland's outlet glaciers. **a**, Helheim Glacier terminates in Sermilik Fjord, southeast Greenland. The glacier is grounded in 600 m of water and has a thick ice mélange. The light coloured lower reaches of the mountains show the glacier's extent before the early 2000s retreat. **b**, Helheim retreated more than 4 km between 2002 and 2005. The retreat was comparable only with that of the 1930s⁵⁸.

polar water even before the mid-1990s, although the Atlantic water layer was probably colder and thinner, given the colder conditions on the shelves. Fjord waters can be renewed ('flushed') through different processes (Fig. 4), including exchanges driven by the shelf circulation, tidal mixing, local winds and estuarine flows^{42,43} (Box 1). How quickly and by what mechanisms the fjords' properties evolved from the pre- to the post-SPNA warming conditions remains poorly understood, owing to the lack of direct velocity measurements. The few that exist suggest fast currents and high-frequency (weekly) variability^{42,43}. The limited moored measurements and repeat surveys, consistent with the fast flows, reveal property changes on sub-seasonal time-scales^{39,43,44}, suggesting that even the large fjords have rapid flushing times, with responses to the proposed shelf warming within months.

Plume dynamics at the ice–ocean interface

A warming of the fjords does not simply translate into an increase in submarine melt. Indeed, the fjords' waters contain much more heat than the glaciers can extract (based on estimated upper bounds for total ice discharge derived from remote sensing⁴⁰). This is because the heat exchange between the ice and the ocean depends on the turbulent exchange across a thin ice–ocean boundary and ultimately on molecular processes that make it fairly inefficient⁴⁵. There are no direct observations for Greenland's glaciers but parallels with Antarctica, theory and models indicate that the boundary layer is dominated by one or more rising, buoyant plume or plumes driven by the release of fresh water from surface melting at the base of the glacier (subglacial discharge) and from submarine melting along the glacier face⁴⁶. The exchange of heat and salt is thus largely dominated by the plumes' properties (temperature, velocity and,

to a lesser extent, salinity⁴⁷) which, in turn, are controlled by a number of oceanic (for example, far field velocity, temperature and salinity) and glaciological (for example, shape of the ice front or surface roughness) parameters in ways that are not fully understood.

One parameter that has emerged as a key factor affecting summer submarine melt rates is the amount and distribution of subglacial discharge. Observations of contrasting winter and summer conditions in the fjords have highlighted the importance of subglacial discharge^{43,44,48}. Additional quantitative modelling studies have shown that an increase in subglacial discharge will lead to an increase in submarine melting^{46,47,49–51}. Other fjord properties that influence submarine melt rates include warmer waters (which increase melt rates), its stratification (the large density difference between Atlantic water and polar water can cause the plume to equilibrate at the Atlantic water–polar water interface, thus limiting the vertical extent of submarine melting^{47,48}), and its circulation (which can both affect the plume and is potentially a source of turbulent kinetic energy for mixing across the ice–ocean boundary).

Proposed mechanisms

From an oceanographic viewpoint, three types of mechanisms are directly implicated in glacier–ocean interactions: thermodynamic processes involved in melting at the calving front, circulation processes that modulate water masses at the calving front, and stress balance ('dynamic') perturbations resulting from the contact between the ice mélange and the calving front.

Submarine discharge and melting

Given the observed changes in the ocean and atmosphere in the mid-1990s there are at least two mechanisms that probably gave rise to increased submarine melt rates (Fig. 5a, b). First, the increase in Atlantic water on the shelves probably resulted in a warmer and thicker layer of Atlantic water in the fjords, both of which will increase submarine melt rates⁴⁷. Second, an increase in subglacial discharge (due to anomalous surface melt⁶) would have increased summer melt rates^{46,47,50,51}. The former may be referred to as melt-driven convection, the latter as convection-driven melting⁴⁶. There are no direct observations of submarine melt rates for Greenland's glaciers (before or after the retreat started). Indirect estimates based on ice-divergence calculations for the Jakobshavn Isbræ are $228 \pm 49 \text{ m yr}^{-1}$ pre-speed up, with an estimated 25% increase due to warming ocean temperatures⁵². Summer estimates based on measuring the oceanic heat flux towards a glacier range from 26 to $1,400 \text{ m yr}^{-1}$ (refs 38, 40, 42, 50), but these are highly uncertain given the unsteady nature of the circulation in the fjord and the assumption that all of the heat flux goes into melting.

Estuarine compared with intermediary fjord circulation

One unresolved issue, relevant to our understanding of how the submarine melt rate may have changed in the mid-1990s, is the extent to which the estuarine-driven circulation (Box 1), due to the release of submarine and surface melt from the glacier, is the dominant 'heat-transporting' circulation in Greenland's glacial fjords^{16,38,47,50,52}. If this were the case, then an increase in subglacial discharge (and hence submarine melting) would increase the transport of heat towards the glacier and potentially further increase the submarine melting. At present, there is no evidence, however, that the estuarine-driven circulation governs the renewal of Atlantic water in the fjords. Numerical simulations have been instrumental in advancing our understanding of the relationship between the plume dynamics, the submarine melt rate and the fjord circulation^{47,50,51,53,54}. However, covering the relevant spatial scales, from millimetres to tens of kilometres, is challenging even for the highest resolution models. Instead, models rely on parameterizations of unresolved processes, some of which are poorly constrained by observations^{46,47}. Furthermore, fjord-scale simulations are sensitive to boundary conditions imposed at the fjord's mouth for which few measurements of variability exist, making it difficult to assess how shelf-driven variability influences submarine melting.

Ice mélange in the fjords

Unlike submarine melting, it is less clear how the warming of the SPNA may have affected the ice mélange — another proposed direct influence on the glaciers²⁷. SPNA warming probably resulted in an increase in the subsurface Atlantic water temperatures in the fjords, but it is unclear what direct impact it had on the surface temperatures in the fjords and, hence, on the ice mélange. However, SPNA warming is highly correlated with an increase in the coastal air temperatures^{16,17}, which, in turn, may affect the structural integrity of the mélange. Changes in sea ice cover outside of the fjord may further affect the mélange.

Glacier retreat during the past century

The recent warming of the upper 1,000 m of the SPNA is unprecedented over the instrumental record of upper ocean temperatures (although a less pronounced warming occurred in the 1960s)^{32,55} (Figs 1b, 3d). During the past century, warming comparable with that of recent decades only occurred in the 1930s as observed from temperature records of the upper 300 m of the North Atlantic⁵⁶; sea-surface temperatures from the eastern subpolar North Atlantic⁵⁷; temperatures (0–40 m depth) from Fylla Bank, west Greenland, 1870 to present⁵⁸ (Fig. 1b); and in the reconstruction of ocean temperatures at the surface and at 300 m from sediment cores in Disko Bay, west Greenland⁵⁹.

Records of glacier frontal position before continuous dedicated satellite radar observations became available (from 1991) are scarce. Cumulative evidence from several studies nevertheless suggests that the only time over the past century when glaciers in southeast and west Greenland retreated as much as in the present day was in the 1930s, consistent with the North Atlantic warming. These include the reconstruction of frontal positions of glaciers in southeast and

west Greenland from photographs (for example, Fig. 2b) and remote sensing^{59–61}, and a reconstruction of calving variability over the past 120 years of one major southeast Greenland glacier from sediment cores⁵⁸. Air temperatures over the ice sheet were also high in the 1930s¹⁴, which, together with ocean warming, would also have led to an increase in submarine melting.

Causes of SPNA warming

The SPNA ocean warming that began in the mid-1990s is manifested as an increase in heat content of the upper 1,000 m (Fig. 3d) and has been associated with a slow down of the subpolar gyre^{55,62}. The warming is attributed to the anomalous inflow of warm, salty, subtropical Atlantic water into the subpolar region⁶³ driven by shifting wind patterns over the North Atlantic^{62,64}. These, in turn, are strongly correlated with the wintertime occurrence of large, quasi-stationary waves in the North Atlantic eddy-driven jet stream (atmospheric blocking) over Greenland and western Europe^{64,65}.

Although progress has been made in explaining SPNA warming, its connection to the large-scale variability of the coupled ocean–atmosphere system remains unclear. Several studies have linked the SPNA changes to the North Atlantic oscillation (NAO)^{66,67}, a dominant mode of atmospheric variability over the North Atlantic (Box 1), which switched from a persistent positive phase in the early 1990s to a negative or quasi-neutral phase until the mid-2000s (Fig. 1c). This inference is consistent with the expected warming of the subpolar and cooling of the subtropical North Atlantic during a negative NAO phase. This opposing behaviour has been used to explain the synchronous and opposite changes in upper ocean heat content anomalies of the subtropical and subpolar gyres from the 1950s to 2000s^{67,68,32} (Fig. 3c, d), and has led investigators to conclude

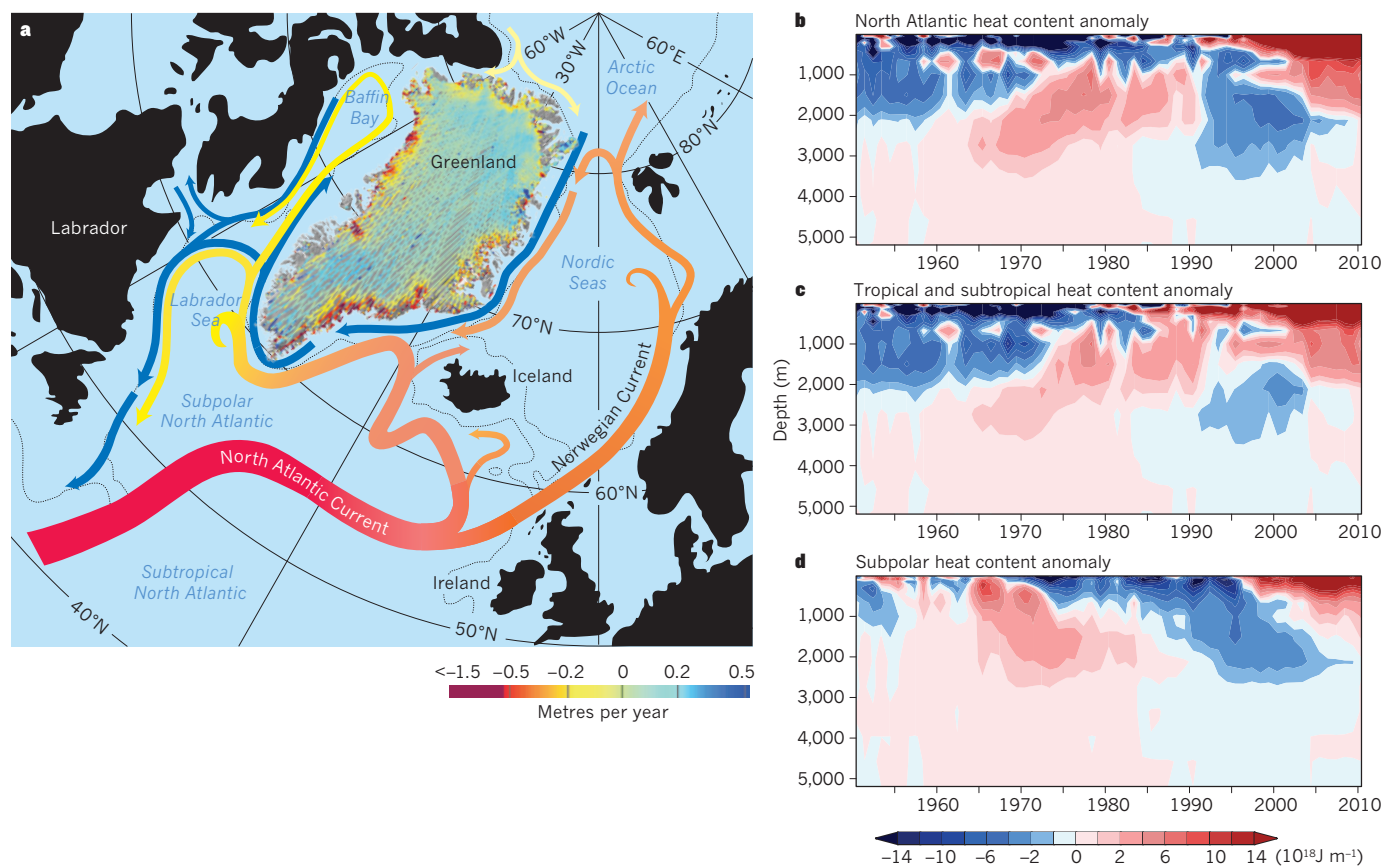


Figure 3 | Thinning of the Greenland ice sheet is concentrated at the margins of the subpolar North Atlantic. **a**, The large-scale ocean circulation around Greenland, indicating the major currents and basins. Atlantic-origin water pathways, red to yellow; Arctic-origin freshwater pathways, blue⁴¹. The dynamic thinning of Greenland is superimposed¹⁹.

b, Heat content anomaly estimates in the North Atlantic as a whole and **c**, separated into tropical and subtropical and **d**, subpolar contributions over the period 1960–2010 (ref. 32). Extremely sparse observational coverage below 700 m depths over much of the period adds significant uncertainties.

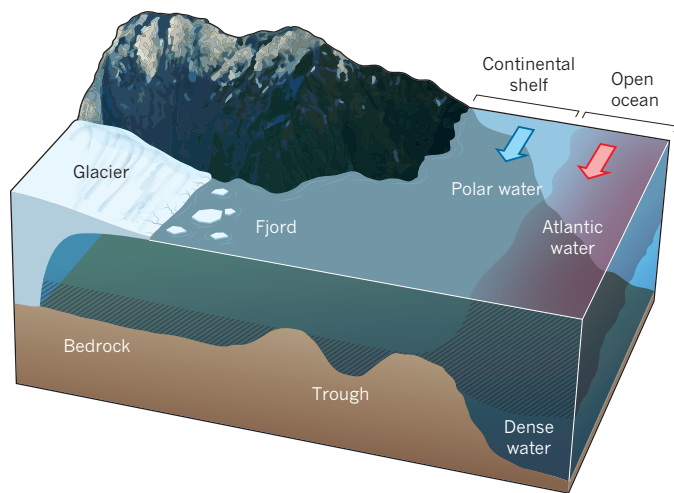


Figure 4 | Fjord and continental shelf exchanges. Warm and salty Atlantic waters (red) of subtropical origin, circulating around the subpolar North Atlantic reach Greenland's glacial fjords at depth (and hence the outlet glaciers) after crossing the continental shelf, where cold, fresh polar waters of Arctic origin flow close to the coast. The ocean-to-glacier link involves a wide range of space and time scales across regions with distinct dynamics.

that Atlantic water variability at Greenland's margins was linked to the NAO²⁸. Recent studies, however, question the role of the NAO as a driver of the recent SPNA variability. In particular, statistical analysis suggests that much of the wind stress, SPNA heat content and sea-surface height anomalies cannot be attributed to the NAO^{64,69}. NAO-driven variability cannot account for the recent simultaneous warming of the upper 1,000–2,000 m of the subpolar and subtropical North Atlantic that started in the late 1990s and has resulted in a large heat content anomaly over the entire North Atlantic (Fig. 3b).

On multidecadal timescales, the warming of the SPNA has a stronger correlation with the Atlantic multidecadal oscillation^{55,56} (AMO), an index^{70–72} that is associated with a range of climate processes including AMOC variability (Box 1). The absolute AMO contains a significant contribution from the global rise in SST⁷³ (Fig. 1c), but a significant correlation between the SPNA SST and the AMO persists even after the global rise in SST is subtracted⁶⁴. In general, neither the sea-surface height variability of the SPNA nor the AMO, both of which are oceanic modes of variability, are connected in any simple way to atmospheric variability⁷⁴.

The emerging picture is that SPNA warming is largely due to increased inflow of warm, subtropical waters into the subpolar region. The subtropical ocean has been accumulating heat⁷⁵ (Fig. 3c), partly due to atmospheric warming³², but it was only after the NAO emerged from its persistent positive phase that some of this heat was transported into the subpolar region. The persistence of wind patterns driving subtropical waters into the subpolar region, a warm AMO phase and the warming trend in the North Atlantic ocean⁷⁵ are all likely contributors to SPNA warming.

The future

The development of skilful predictive capabilities of ice-sheet–ocean interactions around Greenland will require an understanding of the relevant forcing factors and the physical processes that are responsible for Greenland's response, and consideration of the effect of the mass loss on the regional and global climate system. We briefly discuss several key observational and modelling challenges, which will require broad scientific community engagement across disciplines.

Identifying research requirements

The challenge at hand is cross-disciplinary, involving oceanographers, glaciologists, meteorologists, and palaeoclimate and climate scientists. Bringing the corresponding modelling and observational

communities together is imperative and requires international collaboration at the researcher and agency levels. The first steps towards this goal were made in June 2013 through an international workshop on Greenland ice-sheet–ocean interactions in Beverly, Massachusetts, sponsored by the US Climate Variability and Predictability Research Program (US CLIVAR).

The details are subject to ongoing discussion, but an overall two-pronged approach to make progress has emerged²⁶. First, dedicated process studies and field campaigns are needed, in which the available diverse observational assets are pooled, to shed light on the mechanisms described in 'Proposed Mechanisms'. The aim is to move from a qualitative to a quantitative description against which theory and numerical models can be tested, as a prerequisite for developing suitable parameterizations for climate models. One difficulty is whether there exists a 'representative system' to study, or whether different mechanisms require studying different systems. Second, quantitative understanding of the forcing functions in relation to climate variability and change will require the design and maintenance of a long-term observation system at the margins of several strategically located glaciers around Greenland.

Technological innovations are required to allow the collection of observations that are not at present possible; for example, to quantify submarine melting and to understand the dominant controls on calving. A particular concern common to the ocean and ice-sheet modelling communities is the requirement of improved bed maps (outlet glaciers, fjords and continental shelf).

Climate change and GrIS mass loss

Projections of atmospheric circulation changes, including the North Atlantic jet stream characteristics, associated changes in surface temperatures, and implied surface melting on the GrIS are key to inferring magnitudes of source waters that feed subglacial discharge. Evidence for polar amplification (that is, the above-global average increase in Arctic near-surface temperatures) has been found in observations and climate model simulations^{76,77}. This implies that subglacial discharge may increase significantly, due both to increased surface melting and to an extended summer melt season. To what extent polar amplification in the coming century might be, in part, offset around Greenland by a reduction in poleward oceanic heat transport — possibly due to changes in the structure of the AMOC — is at present unclear⁷⁸.

Ancillary to the projection of subglacial discharge are the required modelling capabilities of the surface, englacial and sub-glacial drainage system, as well as the connection between glacial hydrology, geometry, submarine melting and calving^{79–81}. An understanding of these processes remains in its infancy and an important research focus in the coming decade. Enhancing observational capabilities will require technological advances not unlike those made in the design of space missions to conduct autonomous measurements in 'remote' and hostile environments. Model simulations are faced with the challenge of bridging roughly seven orders of magnitudes of scales, and representing different physical processes (ice–ocean boundary layer to North Atlantic basin scale; Fig 4) — they need to solve multiscale and multiphysics problems.

Quantification of future contributions of oceanic heat delivery to the margins of the GrIS is tied to skilful projections of North Atlantic Ocean circulation changes, which are strongly tied to atmospheric circulation changes⁸². In addition, long-term changes in oceanic heat uptake, storage, and transports contribute to a spatio-temporally complex warming pattern around Greenland. The skill of climate model projections remains at present difficult to quantify, and results depend on future emissions scenarios. A suite of climate model projections (CMIP3) from the IPCC Fourth Assessment Report suggest warming of relevant subsurface (200–500 m) water masses that are substantially larger compared with recent warming rates, and which eventually penetrate to the northern margins of the GrIS⁷⁸. The implication is that outlet glaciers in northern Greenland that at present support 'floating ice tongues'

and show little mass loss (Petermann glacier in the northwest and outlet glaciers of the Northeast Greenland Ice Stream) may become more vulnerable to oceanic forcing. Serious limitations of heat content estimates (such as the one in Fig. 2)^{32,75} are their construction from sparse and uneven spatio-temporal sampling of the global ocean. Concerted efforts are required to establish and sustain a global ocean-observing framework that satisfies stringent climate-quality requirements⁸³.

GrIS mass loss and North Atlantic climate

Finally, we shift the focus from how Greenland responds to climate change to what potential impacts the mass loss has on the climate system. On decadal to centennial timescales the two main perceived effects are sea-level change — directly through oceanic mass increase and its spatio-temporal adjustment due to changes in ocean dynamics^{84,85}, and indirectly through glacial isostatic adjustment (Box 1) effects⁸⁶ — and the impact of surface freshening on the AMOC, its associated meridional heat transport and effects on climate^{87–90}.

Very large conceptual discrepancies remain between impact studies of North Atlantic freshening in terms of magnitudes of freshwater fluxes considered (between 0.01 and 1 Sv), input locations (coastally confined compared with spread out over the interior) and model resolutions considered. Simulations with eddy-permitting models (spatial resolutions of 10 to 25 km) show very different response patterns compared with those realized by current generation climate models (about 100 km). However, none of these studies resolve the first baroclinic Rossby radius of deformation (about 7 km), casting doubts as to whether exchange processes between the boundary currents and the interior (in particular through mesoscale eddies) are correctly represented⁹¹. The amount of freshening that reaches the interior convection sites (together with gradual transformation of Atlantic water masses in the boundary currents) may determine the degree to which the North Atlantic circulation responds, its impact on the atmospheric circulation and potential climate shifts over the continents.

Discussions regarding sea-level implications are already available⁹², and so our focus is on mass loss projections. The absence of available coupled climate–ice-sheet models that are able to resolve outlet glacier flow, include accurate ice-flow dynamics and ice physics (in terms of glacial hydrology, calving models, ice–ocean coupling and moving ice–ocean interface), has led to attempts to provide Greenland mass loss estimates, either through consideration of upper bounds on physically feasible ice flow^{13,93}, or lower bounds from observed present-day perturbations⁹⁴, or forced simulations with current-generation ice-sheet models of varying complexity^{95–97}. The range from 0.01 m to 0.54 m of eustatic sea-level rise until 2100 from Greenland ice dynamics reflects the current uncertainties in these projections. It is important to remember that regional sea level is the variable of more direct societal relevance for coastal communities, and which may exceed the global mean considered here by a factor of five⁹¹. A serious limitation to the verification, validation and calibration of ice-sheet simulations is the near-absence of crucial measurements of conditions in the interior and at the bed. Ice-sheet modelling, therefore, represents a grand challenge computational inverse problem.

An inter-generational scientific challenge

Reducing the uncertainty in projected contributions to sea-level rise from Greenland ice dynamics, as well as ascertaining the reliability of estimated upper bounds requires detailed cross-disciplinary process understanding and vastly improved simulation capabilities in all of the aspects discussed in this Review. Such understanding can only come through much expanded, internationally coordinated observational assets, both at the small-scale process level, and of large-scale circulation changes. It involves the design and deployment of new instruments on the ground, at sea and in space; the maintenance of crucial *in-situ* and satellite observing systems; and the collection of

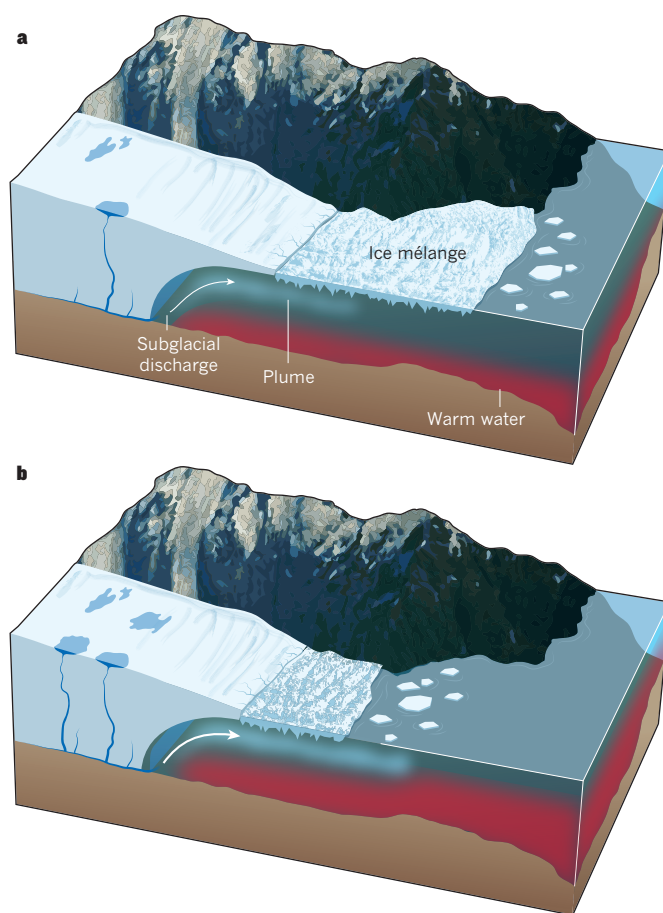


Figure 5 | Submarine melting. Warming subsurface ocean waters and increased glacier surface melt resulted in increased submarine melting and, potentially, a weakened ice mélange at the marine margins of Greenland's outlet glaciers. **a**, Pre-retreat conditions include relatively cold waters, limited subglacial discharge and a thick ice mélange. **b**, Retreat conditions include warm fjord waters, increased subglacial discharge and weakened mélange.

geological records to allow the reconstruction of palaeo-ice-stream evolution through the Holocene. These should be accompanied by rigorous approaches to synthesize the heterogeneous data streams into a coherent dynamic framework⁹⁸. Sustaining such observations over sufficiently long periods to provide records of useful quality for climate research⁹⁹ is a serious inter-generational challenge¹⁰⁰. ■

Received 30 September; accepted 25 October 2013.

1. Shepherd, A. *et al.* A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
2. Hanna, E. *et al.* Ice-sheet mass balance and climate change. *Nature* **498**, 51–59 (2013).
3. Church, J. A. *et al.* Revisiting the Earth's sea-level and energy budgets from 1961 to 2008. *Geophys. Res. Lett.* **38**, L18601 (2011).
4. Dickson, R. *et al.* Current estimates of freshwater flux through Arctic and subarctic seas. *Prog. Oceanogr.* **73**, 210–230 (2007).
5. Bamber, J., van den Broeke, M., Ettema, J., Lenaerts, J. & Rignot, E. Recent large increases in freshwater fluxes from Greenland into the North Atlantic. *Geophys. Res. Lett.* **39**, L19501 (2012).
6. van den Broeke, M. *et al.* Partitioning recent Greenland mass loss. *Science* **326**, 984–986 (2009).
7. Krabill, W. Greenland Ice Sheet: increased coastal thinning. *Geophys. Res. Lett.* **31**, L24402 (2004).
8. Hanna, E. *et al.* Greenland ice sheet surface mass balance 1870 to 2010 based on Twentieth century reanalysis, and links with global climate forcing. *J. Geophys. Res.* **116**, D24121 (2011).
9. Sole, A., Payne, T., Bamber, J., Nienow, P. & Krabill, W. Testing hypotheses of the cause of peripheral thinning of the Greenland ice sheet: is land-terminating ice thinning at anomalously high rates? *Cryosphere* **2**, 205–218 (2008).
10. Rignot, E. & Kanagaratnam, P. Changes in the velocity structure of the Greenland ice sheet. *Science* **311**, 986–990 (2006).

11. Howat, I. M., Joughin, I. & Scambos, T. A. Rapid changes in ice discharge from Greenland outlet glaciers. *Science* **315**, 1559–1561 (2007).
12. Khan, S. A., Wahr, J., Bevis, M., Velicogna, I. & Kendrick, E. Spread of ice mass loss into northwest Greenland observed by GRACE and GPS. *Geophys. Res. Lett.* **37**, L06501 (2010).
13. Moon, T., Joughin, I., Smith, B. & Howat, I. 21st-century evolution of Greenland outlet glacier velocities. *Science* **336**, 576–578 (2012).
14. Box, J. E., Yang, L., Bromwich, D. H. & Bai, L.-S. Greenland ice sheet surface air temperature variability: 1840–2007. *J. Clim.* **22**, 4029–4049 (2009).
15. Bersch, M., Yashayaev, I. & Koltermann, K. P. Recent changes of the thermohaline circulation in the subpolar North Atlantic. *Ocean Dyn.* **57**, 223–235 (2007).
16. Hanna, E. *et al.* The influence of North Atlantic atmospheric and oceanic forcing effects on 1900–2010 Greenland summer climate and ice melt/runoff. *Int. J. Climatol.* **33**, 862–880 (2013).
17. Hall, D. K. *et al.* Variability in the surface temperature and melt extent of the Greenland ice sheet from MODIS. *Geophys. Res. Lett.* **40**, 2114–2120 (2013).
18. Nick, F. M., Vieli, A., Howat, I. M. & Joughin, I. Large-scale changes in Greenland outlet glacier dynamics triggered at the terminus. *Nature Geosci.* **2**, 110–114 (2009).
19. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461**, 971–975 (2009).
20. Joughin, I., Abdalati, W. & Fahnestock, M. Large fluctuations in speed on Greenland's Jakobshavn Isbrae glacier. *Nature* **432**, 608–610 (2004).
21. Thomas, R. H. Force-perturbation analysis of recent thinning and acceleration of Jakobshavn Isbrae, Greenland. *J. Glaciol.* **50**, 57–66 (2004).
22. Luckman, A., Murray, T., de Lange, R. & Hanna, E. Rapid and synchronous ice-dynamic changes in East Greenland. *Geophys. Res. Lett.* **33**, L03503 (2006).
23. Vieli, A. & Nick, F. M. Understanding and modelling rapid dynamic changes of tidewater outlet glaciers: issues and implications. *Surv. Geophys.* **32**, 437–458 (2011).
24. Joughin, I. *et al.* Seasonal to decadal scale variations in the surface velocity of Jakobshavn Isbrae, Greenland: observation and model-based analysis. *J. Geophys. Res.* **117**, F02030 (2012).
25. Joughin, I., Alley, R. & Holland, D. Ice-sheet response to oceanic forcing. *Science* **338**, 1172–1176 (2012).
26. Straneo, F. *et al.* Challenges to understand the dynamic response of Greenland's marine terminating glaciers to oceanic and atmospheric forcing. *Bull. Am. Meteorol. Soc.* **94**, 1131–1144 (2013).
27. Amundson, J. M. *et al.* Ice mélange dynamics and implications for terminus stability, Jakobshavn Isbrae, Greenland. *J. Geophys. Res.* **115**, F01005 (2010).
28. Holland, D. M., Thomas, R. H., de Young, B., Ribergaard, M. H. & Lyberth, B. Acceleration of Jakobshavn Isbrae triggered by warm subsurface ocean waters. *Nature Geosci.* **1**, 659–664 (2008).
29. Yashayaev, I. Hydrographic changes in the Labrador Sea, 1960–2005. *Prog. Oceanogr.* **73**, 242–276 (2007).
30. Våge, K. *et al.* The Irminger Gyre: Circulation, convection, and interannual variability. *Deep Sea Res. Part I* **58**, 590–614 (2011).
31. Sutherland, D. A. & Pickart, R. S. The east Greenland coastal current: structure, variability, and forcing. *Prog. Oceanogr.* **78**, 58–77 (2008).
32. Williams, R. G., Roussinov, V., Smith, D. & Lozier, S. Decadal evolution of ocean thermal anomalies 1 in the North Atlantic: the effect of Ekman, overturning and horizontal transport. *J. Clim.* <http://dx.doi.org/10.1175/JCLI-D-12-00234.1> (2013).
33. Myers, P. G. & Ribergaard, M. H. Warming of the Polar Water in Disko Bay and potential impact on Jakobshavn Isbrae. *J. Phys. Oceanogr.* <http://dx.doi.org/10.1175/JPO-D-12-051.1> (2013).
34. Myers, P. G., Kulan, N. & Ribergaard, M. H. Irminger water variability in the west Greenland current. *Geophys. Res. Lett.* **34**, L17601 (2007).
35. Zweng, M. M. & Münchow, A. Warming and freshening of Baffin Bay, 1916–2003. *J. Geophys. Res.* **111**, C07016 (2006).
36. Sutherland, D. A. *et al.* Atlantic water variability on the Southeast Greenland continental shelf and its relationship to SST and bathymetry. *J. Geophys. Res. Oceans* **118**, 847–855 (2013).
37. Christoffersen, P. *et al.* Warming of waters in an East Greenland fjord prior to glacier retreat: mechanisms and connection to large-scale atmospheric conditions. *Cryosphere* **5**, 701–714 (2011).
38. Rignot, E., Koppes, M. C. & Velicogna, I. Rapid submarine melting of the calving faces of West Greenland glaciers. *Nature Geosci.* **3**, 187–191 (2010).
39. Straneo, F. *et al.* Rapid circulation of warm subtropical waters in a major glacial fjord in East Greenland. *Nature Geosci.* **3**, 182–186 (2010).
40. Johnson, H. L., Münchow, A., Falkner, K. K. & Melling, H. Ocean circulation and properties in Petermann Fjord, Greenland. *J. Geophys. Res.* **116**, C01003 (2011).
41. Straneo, F. *et al.* Characteristics of ocean waters reaching Greenland's glaciers. *Ann. Glaciol.* **53**, 202–210 (2012).
42. Sutherland, D. A. & Straneo, F. Estimating ocean heat transports and submarine melt rates in Sermilik Fjord, Greenland, using lowered acoustic Doppler current profiler (LADCP) velocity profiles. *Ann. Glaciol.* **53**, 50–58 (2012).
43. Mortensen, J. *et al.* On the seasonal freshwater stratification in the proximity of fast-flowing tidewater outlet glaciers in a sub-Arctic sill fjord. *J. Geophys. Res. Oceans* **118**, 1382–1395 (2013).
44. Mortensen, J., Lennert, K., Bendtsen, J. & Rysgaard, S. Heat sources for glacial melt in a sub-Arctic fjord (Godthåbsfjord) in contact with the Greenland Ice Sheet. *J. Geophys. Res.* **116**, C01013 (2011).
45. Holland, D. M. & Jenkins, A. Modeling thermodynamic ice–ocean interactions at the base of an ice shelf. *J. Phys. Oceanogr.* **29**, 1787–1800 (1999).
46. Jenkins, A. Convection-driven melting near the grounding lines of ice shelves and tidewater glaciers. *J. Phys. Oceanogr.* **41**, 2279–2294 (2011).
47. Sciascia, R., Straneo, F., Cenedese, C. & Heimbach, P. Seasonal variability of submarine melt rate and circulation in an East Greenland fjord. *J. Geophys. Res.* **118**, 2492–2506 (2013).
48. Straneo, F. *et al.* Impact of fjord dynamics and glacial runoff on the circulation near Helheim Glacier. *Nature Geosci.* **4**, 322–327 (2011).
49. Motyka, R. J., Hunter, L., Echelmeyer, K. A. & Connor, C. Submarine melting at the terminus of a temperate tidewater glacier, LeConte Glacier, Alaska, USA. *Ann. Glaciol.* **36**, 57–65 (2003).
50. Xu, Y., Rignot, E., Fenty, I., Menemenlis, D. & Flexas, M. M. Subaqueous melting of Store Glacier, West Greenland from three-dimensional, high-resolution numerical modeling and ocean observations. *Geophys. Res. Lett.* **40**, 4648–4653 (2013).
51. Xu, Y., Rignot, E., Menemenlis, D. & Koppes, M. Numerical experiments on subaqueous melting of Greenland tidewater glaciers in response to ocean warming and enhanced subglacial discharge. *Ann. Glaciol.* **53**, 229–234 (2012).
52. Motyka, R. J. *et al.* Submarine melting of the 1985 Jakobshavn Isbrae floating tongue and the triggering of the current retreat. *J. Geophys. Res.* **116**, F01007 (2011).
53. Mugford, R. I. & Dowdeswell, J. A. Modeling glacial meltwater plume dynamics and sedimentation in high-latitude fjords. *J. Geophys. Res.* **116**, F01023 (2011).
54. Salcedo-Castro, J., Bourgault, D. & deYoung, B. Circulation induced by subglacial discharge in glacial fjords results from idealized numerical simulations. *Cont. Shelf Res.* **31**, 1396–1406 (2011).
55. Häkkinen, S., Rhines, P. B. & Worthen, D. L. Northern North Atlantic sea surface height and ocean heat content variability. *J. Geophys. Res. Oceans* **118**, 3670–3678 (2013).
56. Polyakov, I. V. *et al.* Multidecadal variability of North Atlantic temperature and salinity during the twentieth century. *J. Clim.* **18**, 4562–4581 (2005).
57. Reverdin, G. North Atlantic subpolar gyre surface variability (1895–2009). *J. Clim.* **23**, 4571–4584 (2010).
58. Andresen, C. S. *et al.* Rapid response of Helheim glacier in Greenland to climate variability over the past century. *Nature Geosci.* **5**, 37–41 (2012).
59. Lloyd, J. M. *et al.* A 100 year record of ocean temperature control on the stability of Jakobshavn Isbrae, West Greenland. *Geology* **39**, 867–870 (2011).
60. Björk, A. A. *et al.* An aerial view of 80 years of climate-related glacier fluctuations in southeast Greenland. *Nature Geosci.* **5**, 427–432 (2012).
61. Howat, I. M. & Eddy, A. Multi-decadal retreat of Greenland's marine-terminating glaciers. *J. Glaciol.* **57**, 389–396 (2011).
62. Häkkinen, S. & Rhines, P. B. Decline of subpolar North Atlantic circulation during the 1990s. *Science* **304**, 555–559 (2004).
63. Hätún, H., Sandø, A. B., Drange, H., Hansen, B. & Valdimarsson, H. Influence of the Atlantic subpolar gyre on the thermohaline circulation. *Science* **309**, 1841–1844 (2005).
64. Häkkinen, S., Rhines, P. B. & Worthen, D. L. Atmospheric blocking and Atlantic multidecadal ocean variability. *Science* **334**, 655–659 (2011).
65. Woollings, T. & Hoskins, B. Simultaneous Atlantic–Pacific blocking and the Northern annular mode. *Q. J. R. Meteorol. Soc.* **134**, 1635–1646 (2008).
66. Hurrell, J. W. Decadal trends in the North Atlantic oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
67. Visbeck, M. *et al.* in *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, Vol. 134 (eds Hurrell, J. W. *et al.*) 113–145 (AGU, 2003).
68. Lozier, M. S. *et al.* The spatial pattern and mechanisms of heat-content change in the North Atlantic. *Science* **319**, 800–803 (2008).
69. Lohmann, K., Drange, H. & Bentsen, M. A possible mechanism for the strong weakening of the North Atlantic subpolar gyre in the mid-1990s. *Geophys. Res. Lett.* **36**, L15602 (2009).
70. Schlesinger, M. E. & Ramanjany, N. An oscillation in the global climate system of period 65–70 years. *Nature* **367**, 723–726 (1994).
71. Enfield, D. B., Mestas-Nunez, A. M. & Trimble, P. J. The Atlantic multidecadal oscillation and its relationship to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.* **28**, 2077–2080 (2001).
72. Polyakov, I. V., Pnyushkov, A. V. & Timokhov, L. A. Warming of the intermediate Atlantic water of the Arctic Ocean in the 2000s. *J. Clim.* **25**, 8362–8370 (2012).
73. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, L12704 (2006).
74. Chhak, K. C., Moore, A. M. & Milliff, R. F. Stochastic forcing of ocean variability by the North Atlantic oscillation. *J. Phys. Oceanogr.* **39**, 162–184 (2009).
75. Levitus, S. *et al.* World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophys. Res. Lett.* **39**, L10603 (2012).
76. Manabe, S. & Stouffer, R. J. Sensitivity of a global climate model to an increase of CO₂ concentration in the atmosphere. *J. Geophys. Res.* **85**, 5529–5554 (1980).
77. Chylek, P., Folland, C. K., Lesins, G., Dubey, M. K. & Wang, M. Arctic air temperature change amplification and the Atlantic multidecadal oscillation. *Geophys. Res. Lett.* **36**, L14801 (2009).
78. Yin, J. *et al.* Different magnitudes of projected subsurface ocean warming around Greenland and Antarctica. *Nature Geosci.* **4**, 524–528 (2011).
79. Post, A., O'Neil, S., Motyka, R. J. & Streveler, G. A complex relationship between calving glaciers and climate. *Eos Trans. AGU* **92**, 305–306 (2011).
80. O'Leary, M. & Christoffersen, P. Calving on tidewater glaciers amplified by submarine frontal melting. *Cryosphere* **7**, 119–128 (2013).
81. Podrasky, D., Truffer, M., Fahnestock, M., Amundson, J. M., Cassotto, R., & Joughin, I. Outlet glacier response to forcing over hourly to interannual

- timescales, Jakobshavn Isbræ, Greenland. *J. Glaciol.* **58**, 1212–1226 (2012).
82. Woollings, T., Gregory, J. M., Pinto, J. G., Meyers, M. & Brayshaw, D. J. Response of the North Atlantic storm track to climate change shaped by ocean-atmosphere coupling. *Nature Geosci.* **5**, 313–317 (2012).
 83. Lindstrom, E. *et al.* A framework for Ocean Observing <http://dx.doi.org/10.5270/OceanObs09-FOO> (UNESCO, 2012).
 84. Stammer, D. Response of the global ocean to Greenland and Antarctic ice melting. *J. Geophys. Res.* **113**, C06022 (2008).
 85. Lorbacher, K., Marsland, S. J., Church, J. A., Griffies, S. M. & Stammer, D. Rapid barotropic sea level rise from ice sheet melting. *J. Geophys. Res.* **117**, C06003 (2012).
 86. Mitrovica, J. X. *et al.* On the robustness of predictions of sea level fingerprints. *Geophys. J. Int.* **187**, 729–742 (2011).
 87. Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
 88. Marsh, R. *et al.* Short-term impacts of enhanced Greenland freshwater fluxes in an eddy-permitting ocean model. *Ocean Sci.* **6**, 749–760 (2010).
 89. Weijer, W., Maltrud, M. E., Hecht, M. W., Dijkstra, H. A. & Klijhuis, M. A. Response of the Atlantic ocean circulation to Greenland ice sheet melting in a strongly-eddy-permitting ocean model. *Geophys. Res. Lett.* **39**, L09606 (2012).
 90. Hu, A. *et al.* Influence of continental ice retreat on future global climate. *J. Clim.* **26**, 3087–3111 (2013).
 91. Gelderloos, R., Katsman, C. A. & Drijfhout, S. S. Assessing the roles of three eddy types in restratifying the Labrador Sea after deep convection. *J. Phys. Oceanogr.* **41**, 2102–2119 (2011).
 92. Stammer, D., Cazenave, A., Ponte, R. M. & Tamisiea, M. E. Causes for contemporary regional sea level changes. *Annu. Rev. Mar. Sci.* **5**, 21–46 (2013).
 93. Pfeffer, W. T., Harper, J. T. & O'Neel, S. Kinematic constraints on glacier contributions to 21st-century sea-level rise. *Science* **321**, 1340–1343 (2008).
 94. Price, S. F., Payne, A. J., Howat, I. M. & Smith, B. E. Committed sea-level rise for the next century from Greenland ice sheet dynamics during the past decade. *Proc. Natl Acad. Sci. USA* **108**, 8978–8983 (2011).
 95. Gillet-Chaulet, F. *et al.* Greenland ice sheet contribution to sea-level rise from a new-generation ice-sheet model. *Cryosphere* **6**, 1561–1576 (2012).
 96. Nick, F. M. *et al.* Future sea-level rise from Greenland's main outlet glaciers in a warming climate. *Nature* **497**, 235–238 (2013).
 97. Nowicki, S. *et al.* Insights into spatial sensitivities of ice mass response to environmental change from the SeaRISE ice sheet modeling project II: Greenland. *J. Geophys. Res. Earth Surf.* **118**, 1025–1044 (2013).
 98. Wunsch, C. & Heimbach, P. in *Ocean Circulation and Climate: A 21st Century Perspective 2nd edn* (eds Siedler, G., Church, J. Gould, J. & Griffies, S.) 553–579 (Elsevier, 2013).
 99. Wouters, B., Bamber, J. L., van den Broeke, M. R., Lenaerts, J. T. M. & Sasgen, I. Limits in detecting acceleration of ice sheet mass loss due to climate variability. *Nature Geosci.* **6**, 613–616 (2013).
 100. Wunsch, C., Schmitt, R. W. & Baker, D. J. Climate change as an intergenerational problem. *Proc. Natl Acad. Sci. USA* **110**, 4435–4436 (2013).

Acknowledgements Part of the work discussed here benefited from discussions within the US CLIVAR Working Group on Greenland Ice Sheet–Ocean Interactions (GRISO). US CLIVAR and its sponsoring agencies are thanked for supporting a workshop on this subject held in Beverly, Massachusetts, from June 4–7, 2013. P.H. gratefully acknowledges core support through the Estimating the Circulation and Climate of the Oceans (ECCO) project, and supplemental funding from NASA, NSF, DOE and NOAA. F.S. gratefully acknowledges funding from NSF, NASA and WHOI's OCCI.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/17ijtf. Correspondence should be addressed to F.S. (fstraneo@whoi.edu).

Coastal flooding by tropical cyclones and sea-level rise

Jonathan D. Woodruff¹, Jennifer L. Irish² & Suzana J. Camargo³

The future impacts of climate change on landfalling tropical cyclones are unclear. Regardless of this uncertainty, flooding by tropical cyclones will increase as a result of accelerated sea-level rise. Under similar rates of rapid sea-level rise during the early Holocene epoch most low-lying sedimentary coastlines were generally much less resilient to storm impacts. Society must learn to live with a rapidly evolving shoreline that is increasingly prone to flooding from tropical cyclones. These impacts can be mitigated partly with adaptive strategies, which include careful stewardship of sediments and reductions in human-induced land subsidence.

Flooding in the context of future storm variability, sea-level rise and shoreline change is one of the most important issues facing coastal populations today. In the regions they affect, tropical cyclones are often the most damaging storms and, therefore, of primary importance when assessing flood risk. It is clear that coastal populations are becoming more prone to extreme flooding from tropical cyclones¹. There is also growing evidence for a future shift in the average global intensity of tropical cyclones towards stronger storms². Although both of these two points are probably true, most researchers would agree that linking the two in terms of cause and effect is in many ways incorrect.

First, significant uncertainty exists as to how tropical cyclone activity will vary regionally, particularly with respect to landfalling storms. Second, the level of regional tropical cyclone activity is just one of the factors that drives the magnitude and frequency of tropical cyclone flooding. For example, the Western North Pacific has been the most prolific tropical cyclone basin over the instrumental record, both in terms of the overall number of tropical cyclones (30% of global activity) and peak tropical cyclone wind intensities (Fig. 1). However, in recent decades this basin accounted for neither the majority of economic nor human losses from tropical cyclones. These records have been held by two of the least active tropical cyclone basins, the North Atlantic (10% of global tropical cyclone activity) and North Indian Ocean (5% of global tropical cyclone activity), respectively. Since 1970, around 65% of all lives lost as a result of tropical cyclones occurred within the North Indian Ocean — equivalent to more than half a million deaths³. Over this same period, more than 60% of all economic losses from tropical cyclones took place in the North Atlantic — amounting to around US\$400 billion³.

Although tropical cyclone activity is relatively low in the North Indian Ocean and the North Atlantic, the frequency of coastal flooding is not. Extreme flooding is prevalent mainly on low-gradient shores, including barrier and deltaic systems; these areas have often also attracted the development of dense population centres. Low-lying coasts are typically composed of soft sediments and are particularly dynamic, with geometries that greatly enhance storm impacts. For these evolving shores, storms provide the dominant mechanism of extreme flooding and erosion — although in this Review we discuss how it is often sea-level rise (SLR) that is the underlying cause of both increasing rates of long-term shoreline retreat and flood frequency. Human factors are of equal importance in terms of influencing coastal impacts by tropical cyclones^{1,4,5}, but this topic is beyond the scope of this Review. However,

at the root of these human factors is the flood-prone landscape on which coastal populations have developed. In these settings, joint consideration of tropical cyclone climatology, relative SLR and shoreline change is crucial for accurate assessments of future flood risks. We focus this Review on these three physical factors, highlighting that rising sea levels will become a dominant driver of increased tropical cyclone flooding irrespective of changes in tropical cyclone activity. We point to population centres most at risk of tropical cyclone impacts — those that are mainly located along dynamic and subsiding sedimentary coasts that will serve to further enhance the impact of future tropical cyclone floods. Finally, we discuss managing risk in the context of an almost certain increase in tropical cyclone flood frequency, and the importance of using a holistic approach to manage coastal systems.

Tropical cyclone climatology

On average, about 90 tropical cyclones occur worldwide per year, with the annual distribution of these events varying among the various tropical cyclone basins⁶. Only about one-fifth of tropical cyclones make landfall with the intensity of a hurricane (defined by wind speeds $\geq 33 \text{ ms}^{-1}$), but coastal impacts by tropical cyclones are due largely to this important subset of storms⁷. Accumulated cyclone energy (ACE) is a common metric for comparing the overall tropical cyclone activity of different tropical cyclone regions; it is calculated by taking the sum of each tropical cyclone's maximum wind speed squared for all storms passing through a selected area. Storm surge is also related to wind speed squared (discussed later), thus ACE is a useful measure of both tropical cyclone activity and tropical cyclone surge potential, all else being equal (for example, ignoring the configuration of a coastline and bathymetry). Spatial variability in ACE highlights anomalously high levels of tropical cyclone activity in the North Pacific, relative to the substantially lower levels of activity within the other tropical cyclone regions (Fig. 1a).

Environmental influences

A warm upper ocean, represented by sea surface temperature (SST), is one of the requirements for tropical cyclone formation and intensification, as is evident by the modulation of tropical cyclone activity in response to the seasons⁸. All else being equal, SST directly relates to the theoretical maximum wind speed that tropical cyclones can reach under specific local environmental conditions⁹. This theoretical maximum wind speed, or potential intensity (PI), is also inversely related to

¹Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA. ²Civil and Environmental Engineering, Virginia Tech, Blacksburg 24061, Virginia, USA. ³Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA.

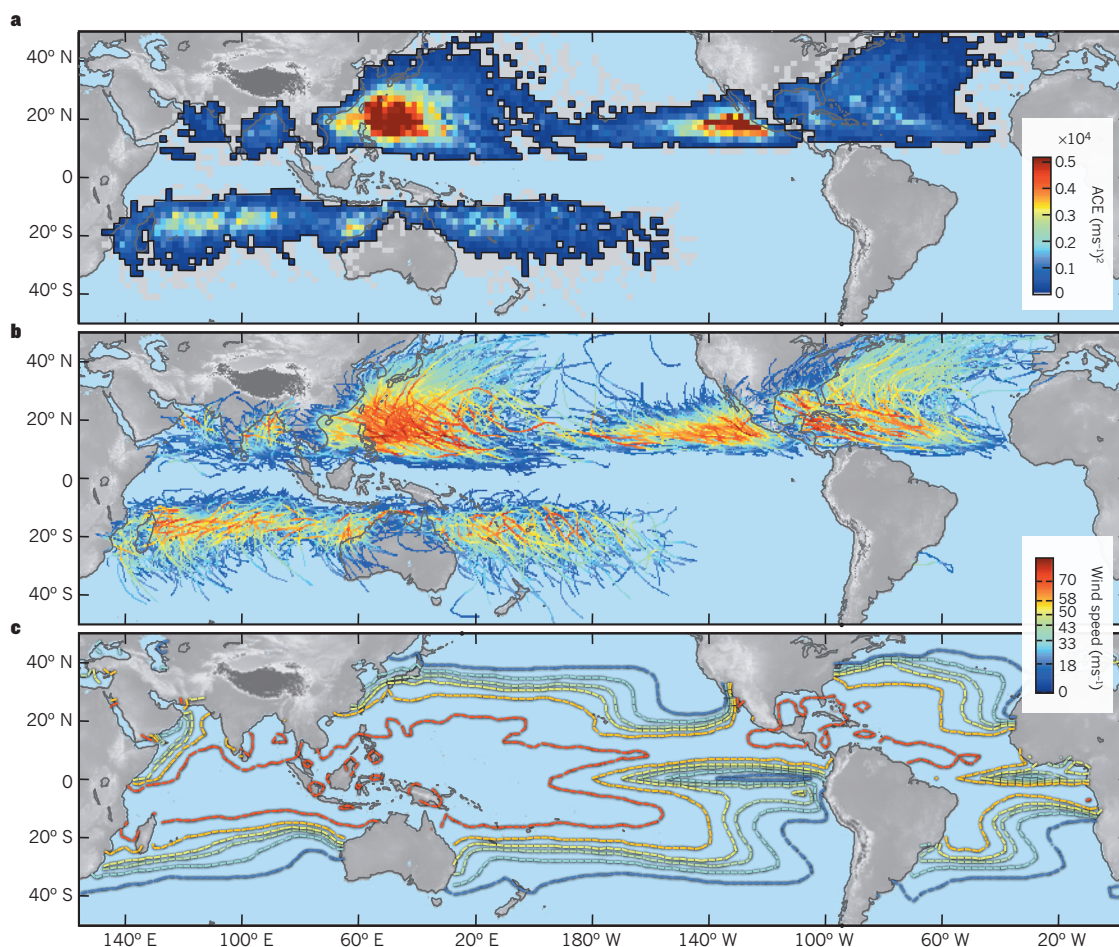


Figure 1 | Global tropical cyclone activity for the period 1981–2010. **a**, Accumulated cyclone energy (ACE). In the Northern Hemisphere, ACE is highest in the western and eastern North Pacific, with lower values in the North Atlantic and Indian Oceans. In the Southern Hemisphere, ACE is highest in the South Indian Ocean. **b**, Historical tropical cyclone tracks. Tracks of intense tropical cyclones concentrate in the western and eastern North Pacific regions, with fewer occurring in the North Atlantic and Southern Hemisphere. Colour scale refers to intensities of tropical cyclone tracks. **c**, Potential intensity for the western North Atlantic and eastern

North Pacific⁸⁷, western North and South Pacific and Indian Ocean⁸⁸, and South Atlantic⁸⁹. Colour scale is the same as in **b** and refers to potential intensity wind speed contours. In the North Atlantic and eastern North Pacific, tropical cyclones with maximum 1 minute sustained wind speeds in excess of 33 ms^{-1} are classified as hurricanes, whereas in the western North Pacific storms meeting this same criterion are called typhoons, and in the Southern Hemisphere they are called severe tropical cyclones. Hurricanes with wind speeds in excess of 50 ms^{-1} are defined as major hurricanes (Categories 3–5).

the outflow temperature where rising air exits a tropical cyclone. The difference between the observed distribution and intensity of tropical cyclone activity (Fig. 1a, b) and PI (Fig. 1c) is due to other environmental factors that are also important in determining tropical cyclone frequency^{10,11}. For example, in the South Atlantic tropical cyclones are scarce (Fig. 1b), despite having a relatively high PI (Fig. 1c). Wind speed in the South Atlantic varies greatly with height in the troposphere (high values of vertical wind shear), which is one important reason for tropical cyclone scarcity in the basin. High vertical wind shear is also a central mechanism for inhibiting tropical cyclone frequency and intensity in the other tropical cyclone regions^{10,12,13}. The amount of humidity in the atmosphere and the presence of pre-existing disturbances, in the form of atmospheric waves and storms that are precursors for tropical cyclone formation also have an important influence on tropical cyclone frequency. All of these factors should be taken into account in future tropical cyclone projections.

Future projections

At the end of the twenty-first century there will probably be fewer, but stronger, storms globally². However, the magnitude range for these predicted changes is still wide, because the different models used to make these projections exhibit different sensitivities to climate change. For

example, projections for changes in the number of tropical cyclones range from –6 to –34% globally, with increases in mean tropical cyclone global wind speed ranging between 2 to 11% by the end of the twenty-first century². Significantly greater uncertainty exists with respect to how tropical cyclone activity will vary regionally, with projected changes up to $\pm 50\%$ in the number of tropical cyclones in individual tropical cyclone basins by the end of the twenty-first century². Similarly, not all ocean basins may experience an increase in tropical cyclone intensity. A statistical downscaling of the tropical cyclone projections of the Coupled Model Intercomparison Project Phase 5 (CMIP5) shows a probable increase in tropical cyclone frequency in the first half of the twenty-first century in the North Atlantic, but the trends in North Atlantic tropical cyclone frequency by the end of the twenty-first century are still uncertain¹⁴. By contrast, North Atlantic tropical cyclone intensity is projected to increase in all climate scenarios by the end of the twenty-first century¹⁵.

Modes of climate variability such as the El Niño–Southern Oscillation (ENSO) and the Madden–Julian Oscillation (MJO) can have a strong regional influence on tropical cyclone frequency and intensity^{10,12,16,17}, and current uncertainties in ENSO and MJO have contributed to the difficulties in obtaining robust global-specific and basin-specific projections^{18,19}. A number of other natural climate modes on various

BOX 1

Tropical cyclone probabilities

Evaluating changes related to tropical cyclone impacts requires a general understanding of the statistical metrics conventionally used for their assessment. The likelihood of winds or flood levels exceeding a threshold is often presented either as the probability of occurrence in a particular year, or with a return period equal to the inverse of this annual probability. For example, a 1% probability of winds or floodwaters exceeding a certain level in any year is equivalent to the event having a 100-year return period. Another useful metric of hazard exposure is the chance that a certain extreme event will be exceeded over a specified interval of time:

$$R = 1 - (1 - Q)^T$$

where R is the chance of an event with an annual exceedance probability of Q occurring over the time period T^{99} . This relationship reveals that there is a 63% chance of a 100-year event ($Q = 1.0\%$) occurring in the next 100 years, and a 10% chance of a 1,000-year event ($Q = 0.1\%$) occurring in the next 100 years. This 10% is still fairly high and serves to highlight why coastal planners often consider events with return periods well beyond the time frame of interest, particularly with respect to sensitive infrastructure. However, the probability of these low-frequency events are the most difficult to constrain, particularly in the context of changes to tropical cyclone climatology.

timescales also influence tropical cyclones in different regions²⁰, and are a source of additional uncertainty. Furthermore, future changes in hybrid storm frequency, including tropical cyclones that undergo extratropical transition, such as Hurricane Sandy in 2012, are largely unknown²¹.

Currently, modes of climate variability, including ENSO and MJO, explain roughly 30–45% of tropical cyclone activity variance within the instrumental historical record⁶. The percentage is much less, however, when considering only storms that make landfall. Furthermore, although these modes of climate variability modulate landfall probabilities in large regions, exact landfall locations are determined by storm tracks, and there is significant variability in tracks both season-to-season and within a single season²². Landfall probabilities are often described as a stochastic process given the high uncertainty associated with local tropical cyclone activity, particularly on shorter timescales (Box 1).

Sea-level rise and tropical cyclone flooding

Global sea level is expected to rise in the upcoming centuries, with a mean global increase that could approach or exceed 1 m by 2100 (ref. 23). SLR is also expected to continue to accelerate through the twenty-first century. Relative SLR at individual sites will vary from this global average²⁴; however, in general, densely populated regions affected by coastal flooding from tropical cyclones have experienced a rate of SLR near or greater than the global average over the instrumental record (Fig. 2).

Before the satellite era, instrumental records of SLR are mostly derived from tide gauges, which record long-term sea-level trends, as well as the sudden rise in water level associated with storm events. Analyses of these time series indicate an increase in extreme high water levels worldwide since 1970, with this increase due almost exclusively to SLR rather than changes in storm climatology²⁵. Longer tide-gauge records along the East Coast of the United States reveal similar results²⁶. However, tide gauge data alone is generally too short to obtain meaningful extreme value statistics²⁷, with derived probabilities that do not account for future, potentially higher, magnitude changes in both sea-level and tropical cyclone activity.

Controls on flooding

Storm surge induced by tropical cyclones depends greatly on coastal geometries, including topography, local shoreline configurations and depth, and individual tropical cyclone characteristics — predominantly the wind speed, storm size and landfall location. The storm's forward motion, angle of approach, and atmospheric pressure drop also influence surge generation. Tidal range and storm timing with the tide; the increase in water level, owing to the presence and local behaviour of shoaling waves; and river discharge and rainfall-driven runoff also

contribute to flooding. However, in coastal regions that experience the most extreme tropical cyclone flooding, the greatest elevated water levels are largely due to wind-driven storm surge. Using a linearized momentum conservation argument, for which bottom friction and other external forces are neglected, it can be shown that wind surge is proportional to:

$$U^2 \frac{W}{h} \quad (1)$$

where U is wind speed, W is the distance over which the wind blows in the same direction, and h is the mean depth over the region where the wind blows²⁸. As equation (1) indicates, wind-driven surge is mainly generated in relatively shallow depths, and where shallow waters extend far offshore. Thus, areas with a relatively broad and shallow continental shelf, such as the western North Atlantic, generally have larger wind-driven surge than areas where offshore slopes are steep, such as the mountainous islands of the western North Pacific and the Caribbean (Fig. 3). However, deltaic low-lying coasts along otherwise steep, less habitable terrain are also particularly susceptible to enhanced tropical cyclone flooding — for example, many of the large population centres in the Bay of Bengal, and sites of growing vulnerability in the western North Pacific^{29,30} (Fig. 3b).

Equation (1) also shows that storm surge is expected to increase with the square of tropical cyclone wind speed. As an example, if tropical cyclone wind intensity for a given tropical cyclone increases by 4% for each degree Celsius of SST warming^{31,32}, from equation (1) we can expect wind surge to increase by 8% for each degree Celsius of SST warming. Damage from storm winds is related to the wind speed cubed, thus compounding impacts related to warming SST³³. However, the approximation for tropical cyclone intensification as a function of warming SST neglects key meteorological influences, which have been discussed previously, including humidity, winds and atmospheric temperature.

Future projections

Coastal flooding probability associated with landfalling tropical cyclones depends both on the probability of tropical cyclone occurrence and the behaviour of relative sea level. Accurate predictions of future flood risk, therefore, must consider the two jointly. The specific role of SLR and the potentially higher occurrence of intense storms in future tropical cyclone flooding have been the focus of a number of recent studies^{34–39}. Many studies assume that tropical cyclone surge and SLR are independent, thus the two may be linearly summed: flood elevation equals surge plus SLR. This approach is a relatively simplistic means of obtaining a global forecast of changes in extreme flood probabilities and associated risk to coastal populations^{34,35}. Although SLR

rates, storm intensification, and time periods differ among studies, the general consensus is for an increase in future extreme flood elevations.

More sophisticated techniques that include a hydrodynamic modelling component directly consider non-linearities between SLR and storm surge^{36,37,39,40}. Simulations in surge-prone Bangladesh were among the first numerical studies to consider both SLR and a potential increase in the tropical cyclone occurrence³⁶. Results show that projected SLR by the 2050s, along with the increased occurrence of intense storms, may inundate up to 15% of the country and could result in a 12% rise in water levels by extreme events. In a more recent study along the coastline of Cairns, Australia, the 100-year return period of a flood event was decreased to a 40-year event using statistically generated storms for the 2050s, along with 0.2 m of SLR and a 10% increase in storm wind speeds³⁷. To assess the combined impact of SLR and changes in tropical cyclone activity for the Atlantic basin a modified joint probability method has been proposed³⁸. For the fourth Intergovernmental Panel on Climate Change (IPCC) “middle-of-the-road” scenario (A1B) on an idealized coast, this study projects the present-day 100-year return period flood elevation becoming the 60-year event by the 2050s. All of the above mentioned results are for relatively moderate rates of SLR by the 2050s and do not account for the more rapid rates of SLR projected for the latter half of the twenty-first century (Fig. 2).

Enhanced rates of relative SLR in regions of rapid land subsidence will further amplify tropical cyclone flooding. This enhanced subsidence is common along populated deltaic and coastal plain systems owing to groundwater, oil and gas extraction, and reductions in fluvial sediment supply. Megacities where past rates of human-induced subsidence exceeded an average of 1 cm yr⁻¹ include Osaka, Japan (2.8 m of subsidence between 1935 and 1995); Manila, Philippines (>1 m of subsidence between 1991 and 2003); Tainjin, China (3.1 m of subsidence between 1959 and 2003); and Tokyo, Japan (5 m of subsidence between 1930 and 1995)^{41,42}. Shanghai, China, is one of the largest megacities that could potentially be affected by elevated rates of relative SLR (2.8 m of subsidence between 1921 and 1995)⁴². Here a 4.3 m projected rise in sea level due to additional land subsidence along the Yangtze River delta by 2100 would result in half of Shanghai being flooded by extreme storm-water levels⁴³. Similar increases in tropical cyclone impacts are projected at other locations where SLR rates are expected to significantly exceed the global average — for example the Red River Delta, Vietnam⁴⁴, and the Mississippi Delta⁴⁵. All of these conclusions assume that no counter-measures are taken to alleviate artificial causes of land subsidence.

One of the most comprehensive projection studies of the combined influence of recent SLR projections and future tropical cyclone climate on storm surge assesses changes in flood probabilities in the New York City region at the end of the twenty-first century³⁹. In this study, a nested modelling technique was used, combining output from global climate model simulations with a physical tropical cyclone model to generate synthetic tropical cyclone tracks for driving hydrodynamic storm surge simulations. Results differ greatly depending on the climate model used, with changes in the return frequency of tropical storms in the New York region ranging from –15% to 290% by the end of the twenty-first century. However, all simulations show increased flooding when a 1 m rise in sea level is included, with the present-day 100-year return period flood event reduced to the 3–20 year event (Box 2 discusses SLR and flooding by Hurricane Sandy in 2012).

These studies highlight current uncertainties associated with future changes in flood frequency that are linked with variability of landfalling tropical cyclones. More importantly, however, they all point to the clear increase in flood frequency associated with an accelerating SLR, regardless of tropical cyclone climatology projections.

Shoreline dynamics

Recent results highlight the importance of relative SLR in increasing coastal flood frequency^{34,35–39}. However, the compounding effects of future shoreline change are not accounted for in most of these assessments. Potential changes in tidal regime may also be important⁴⁶. Coastlines vary greatly in their morphology; however, broad low-lying regions at the greatest risk of tropical cyclone flooding generally share the commonality of being fairly dynamic (Fig. 3). These low-lying shores are often built by mobile sediments (for example, barrier beaches and deltaic coastlines) and/or by biogenic systems (for example, reefs, mangrove wetlands and salt marshes) that are particularly susceptible to climatic and anthropogenic stressors^{47–49}. The frequency and intensity of tropical cyclone flooding has been, and will continue to be, tightly coupled to the morphological development of these coastal systems.

Geomorphic function of tropical cyclones

Tropical cyclones are natural phenomena that have greatly contributed to the morphology of modern shorelines. In many cases, storms serve as a construction mechanism. For instance, sands along the back of barrier beaches are largely storm derived. Deposits from sediments overwashing barrier islands might provide a key mechanism for determining

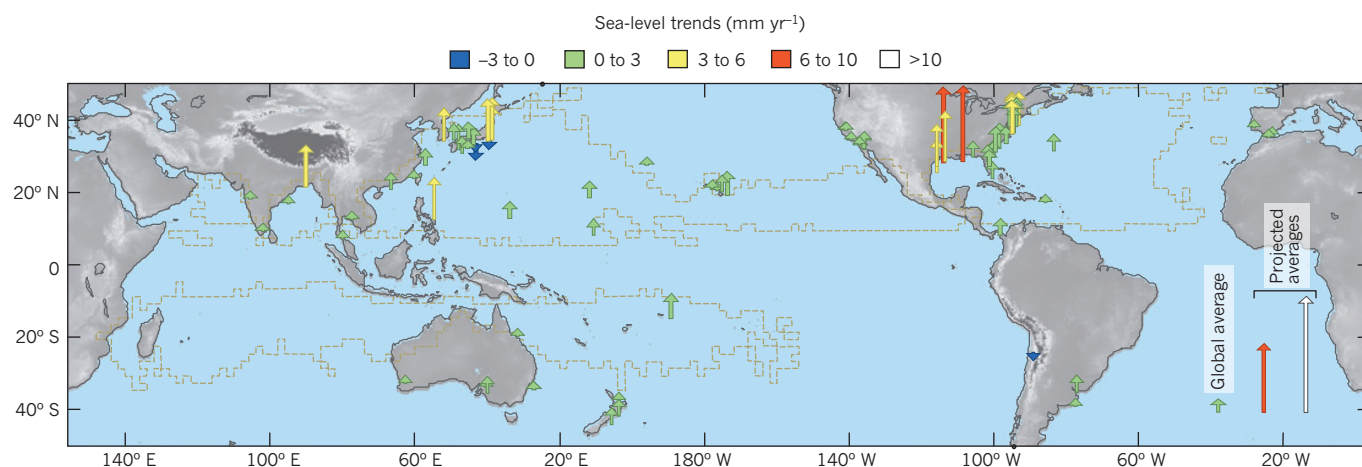


Figure 2 | Global sea-level trends. Local sea-level trends based on individual tidal gauge records more than 50 years old^{24,90}. Green arrows indicate regions where rates of SLR have been near the long-term global average, whereas red and yellow indicate areas where SLR exceeds the global mean. For comparison, arrows on the bottom right show (from left to right) the global instrumental averages from 1900 to present, the projected average rate from present to 2100, and the projected rate at 2100 (ref. 23; see Fig. 4b for SLR time series derived from ref. 23). Dashed lines outline regions of tropical cyclone activity defined

by ACE in Fig. 1a. Spatial coverage is limited by the availability of long-term tide gauge records. However, most of the key population centres affected by tropical cyclones are focused in locations of rising sea level. For instance, by 2020, of the world's top 30 megacities 13 are projected to be along coasts affected by tropical cyclones⁹¹ (see Fig. 3 for locations). With the exception of Chennai, India, all of these population centres have experienced a rise in relative sea level in recent decades, with rates at 10 of these 13 locations greater than the global mean^{41,90,92,93}. Figure adapted with permission from ref. 94.

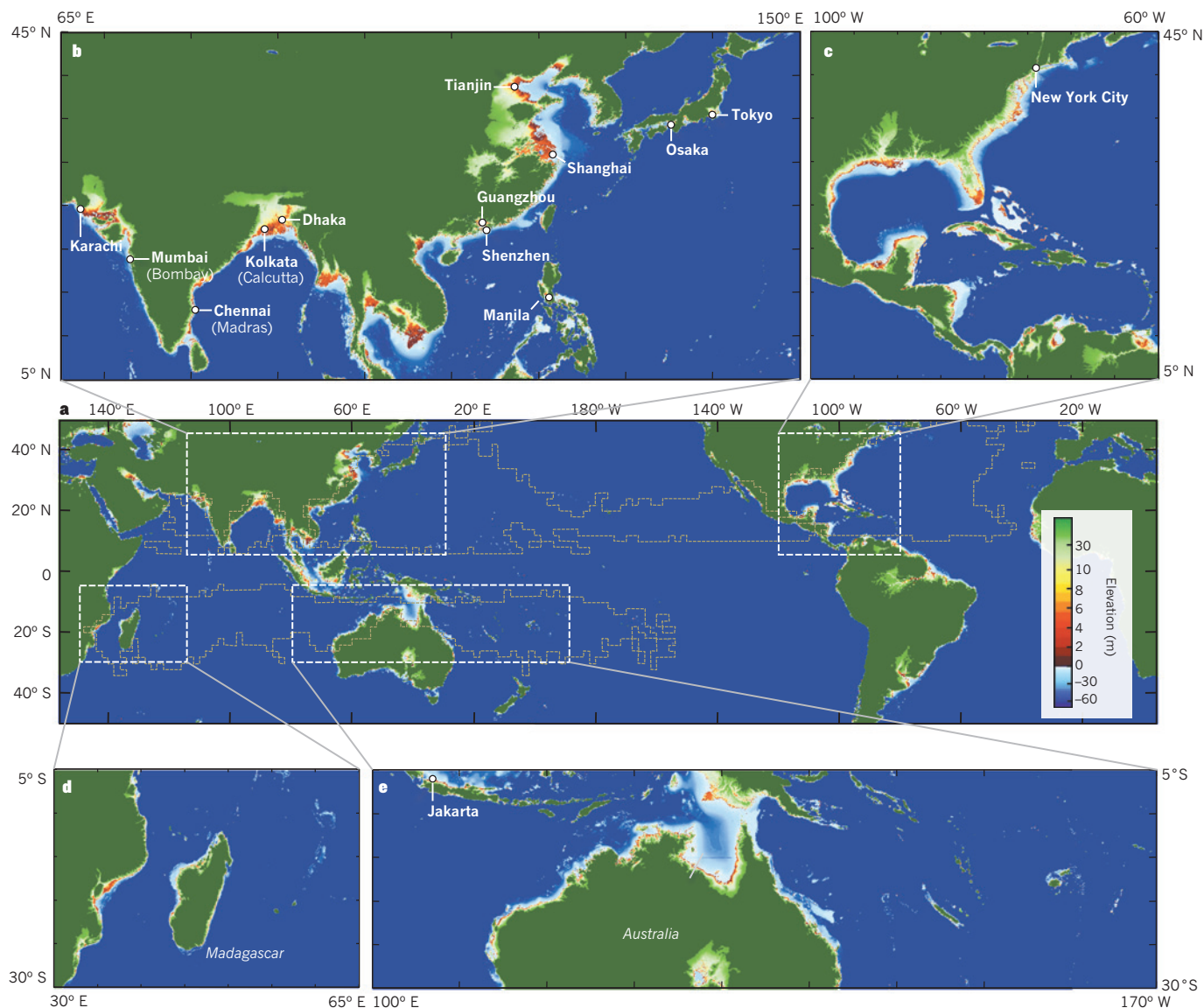


Figure 3 | Coastlines with broad low-lying elevations and shallow abutting bathymetry. **a**, Regions where storm surge is enhanced by shallow depths offshore are shown in pale blue, and low-lying regions generally at a greater risk of coastal flooding are shown in red. Regions of tropical cyclone activity defined by ACE (Fig. 1a) are outlined by grey dashed lines in **a**. Broad regions of low-lying topography and shallow near-shore bathymetry are a fairly good proxy for dynamic and evolving low-gradient shorelines. **b**, The expansive low-lying regions in the Western North Pacific and North Indian Ocean are mainly along deltaic systems that are composed of unconsolidated subsiding

sediments. **c**, Similarly, most of the low-lying coasts affected by tropical cyclones in the Gulf of Mexico and the Western North Atlantic are composed of soft sediments often fronted by dynamic barrier beach systems. Finally, small-island nations affected by tropical cyclones, often identified in **b–e** as isolated light blue regions, are typically fronted by living reef and mangrove systems, which are particularly sensitive to changing environmental conditions. Topographic and bathymetric data are from ref. 95. Coastal cities indicated with circles are ranked among the top 30 of the world's largest urban centres by 2025 (ref. 91).

vertical accretion rates within back-barrier marshes^{50–52}. Furthermore, waves from distant tropical cyclones frequently mobilize offshore sediments that are normally unavailable for littoral transport, allowing this material to be redistributed along the shore face and shallow shelf⁵³. Storms are also largely responsible for sediment redistribution across barrier reef systems⁵⁴, as well as the building of successive beach ridges along seaward advancing or prograding coastlines (commonly referred to as beach ridge plains)⁵⁵.

Naturally, tropical cyclones also erode shorelines, and the building of back-barrier environments often occurs at the expense of an eroding foreshore^{49,56}. Ultimately, this net transport of sediment from the foreshore to the backshore results in the landward retreat of the entire barrier beach system through a barrier rollover mechanism⁴⁹. Mechanisms of shoreline retreat can be complex, with rates governed not only by SLR,

but also by sediment supply and the coastline's pre-existing configuration and lithology (geological or glacial inheritance). The opening of new inlets by storms can also be particularly destructive to barriers because this is often where the greatest loss of beach sediment is observed^{57–59}. Newly formed tidal inlet deltas act as a significant sink for beach sediments. The opening of new inlets can also measurably change tidal exchange and allow ocean surges to more effectively propagate inland. Thus, the surge hazard will be significantly greater if a tropical cyclone occurs while a new inlet remains open. Similarly, inland areas become more vulnerable to tropical cyclone surges through barrier island degradation and inlet formation. Although wide and high barrier islands serve as a natural surge impediment, degraded narrow, low barrier islands readily allow overwash and breaching during tropical cyclones, leading to increased surge levels behind these coastal barriers⁶⁰.

Potentially, many of these coastal systems have tipping points, at which coupled changes in SLR, vegetation coverage and sediment supply result in rapid conversion from one equilibrium state to another, for example gradual barrier island migration compared with complete break up of the barrier island system⁴⁹, or salt marshes compared with open-water tidal flats^{61,62}. Furthermore, the landward retreat of inhabited barrier beaches is inhibited by artificial structures, resulting in shoreline degradation and a loss of the natural buffer that protects infrastructure and homes from large wave forces during tropical cyclone events.

Although initial damage to coastal landforms by tropical cyclones often seems catastrophic, given enough time, these coastal systems generally have the means to recover. The entire barrier beach profile can rebuild if there is sufficient sediment supply^{57,63}, storm-produced inlets can close, and vegetative cover and reef systems can regrow⁶⁴. Shoreline resilience to severe tropical cyclone disturbance requires that enough time lapses between extreme events to allow for recovery; barrier^{57,65} and reef systems⁶⁶ are particularly vulnerable to subsequent flood events during this recovery period.

Tropical cyclone climatology partly drives the length of recovery time that coastal systems have between storm disruptions. However, extreme-value flood statistics consistently point towards SLR as a competing, if not more important, factor in driving the frequency of extreme coastal flooding by tropical cyclones. Thus, although storms provide the dominant mechanism for erosion, it is often an increase in SLR and/or a drop in sediment supply that is the true underlying cause of long-term rates of shoreline retreat⁶³.

Insight from Holocene shoreline development.

Global SLR rates during the early Holocene (roughly 11,500 to 7,000 years before present), are of the same order as many current projections of global SLR by the end of the twenty-first century, about 1 cm yr⁻¹ (Fig. 4). The form and behaviour of shorelines during this earlier period of rapid SLR therefore serves as an important analogue of future shoreline change (although differences exist, including the location of the coast and sediment availability). Often, SLR during this time period was too fast for landforms such as barrier beaches to remain stable, resulting in submergence or rapid landward retreat of these systems⁶⁷. Remnants of relic back-barrier salt marsh and estuarine material are observed kilometres offshore and are evidence of substantial shoreline retreat during the early to mid-Holocene^{68,69}. This period of wide-spread shoreline instability is commonly referred to as the Holocene transgression, a

period of rapid landward retreat of many low-lying sedimentary coastlines in response to high rates of SLR.

In general, the configuration and current function of most modern low-gradient shorelines only established themselves after a significant decline in global SLR rates, beginning around 9,000 to 6,000 years ago (Fig. 4). Rates of sea-level change for the next 6,000 years or so vary regionally⁷⁰, from areas of little change to areas of both net SLR and net sea-level fall. However, with the exception of regions of significant tectonic activity or rapid isostatic adjustment, most coastlines affected by tropical cyclones have experienced moderate rates of sea-level change over the past few millennia relative to the rapid SLR rates of the early Holocene. Examples of current coastal settings, for which the existing forms and behaviours commonly established themselves under these fairly modest rates of sea-level change, include most of the world's deltaic systems⁷¹, barrier beaches^{67,72,73}, contemporary beach ridge and chenier plains^{74,75}, wetland marshes^{76,77} and mangrove wetlands^{78,79}.

Although rates of sea-level change remained relatively low over the later Holocene, tropical cyclone activity did not (Fig. 4c). Statistically significant intervals of both quiescence and increased tropical cyclone activity are evident in the timing of coarse-grained, tropical-cyclone-induced event deposits from back-barrier salt marshes and coastal ponds^{80,81}. Overwash deposits can be delineated within these back-barrier environments because they are later covered by finer-grained organic substrate once sheltered conditions resume. These palaeo-storm records, therefore, not only provide evidence of changes in storm activity over the past few millennia, but they also point to the resilience of barrier beach systems to storms during times of modest sea-level change. By contrast, there is a lack of early Holocene storm deposits preserved behind the modern coast, which points to the seaward location of past shorelines and the frequent reworking of back-barrier sediments by rapid shoreline retreat when past global rates of SLR were elevated to the levels projected for the end of the twenty-first century (Fig. 4).

Storm-induced beach ridges in the South Pacific and South Indian Oceans also serve as a reliable marker of tropical cyclone activity, supplementing overwash deposit information from the North Atlantic and North Pacific⁸² (Fig. 4c). These beach ridge tropical-cyclone-proxies are preserved along shorelines that have been prograding, partly due to moderate rates of sea-level fall over the past 6,000–7,000 years⁸³. Similar to back-barrier overwash reconstructions, the onset for the formation of these beach-ridge shorelines begins only after the Holocene transgression. These shorelines were either stationary or retreating landward before this interval, because of rapid rates of relative SLR.

BOX 2

Sea-level rise and Hurricane Sandy

On October 29, 2012 Hurricane Sandy inundated New York City at high tide, raising water levels to 3.5 m above mean sea level at the Battery (located at the south end of Manhattan Island). Historical records indicate that this event may have exceeded the maximum water levels of the previous highest recorded flood, during a hurricane in 1821 when the water rose roughly 3.2 m above mean sea level at the time¹⁰⁰. However, the 1821 event occurred closer to low tide and when mean sea level at the Battery was roughly 0.5 m lower than present⁹⁴. If the 1821 event were to occur at today's higher sea level and at high tide the resulting flood level for the event would probably have exceeded that observed during Hurricane Sandy. Thus, although Sandy was potentially record-breaking in terms of the overall water elevation reached, it was certainly not unique in terms of its overall surge, with sea-level rise and tides two of the primary causes of Sandy's very high water levels relative the 1821 hurricane event. Flooding as a result of Hurricane Sandy is shown here along the New Jersey coast.



MASTER SGT. MARK OLSEN, US AIR FORCE

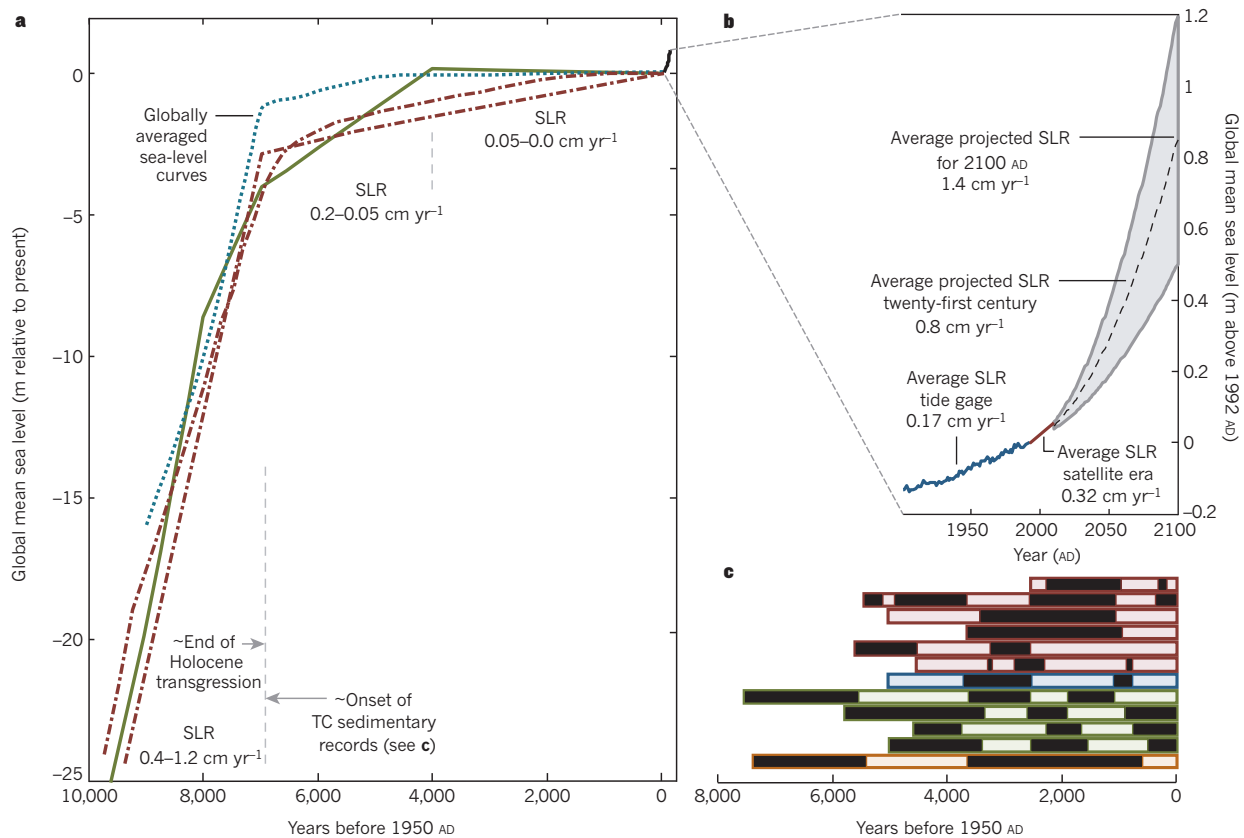


Figure 4 | Mean global sea level along with patterns and extent of preserved sedimentary records of tropical cyclone activity following the most recent glacial maximum. **a**, Four separate estimates of global sea-level elevation since 10,000 years before present^{96–98}, with **b**, associated SLR observed over the twentieth century²³. The twenty-first century projections between intermediate high (IH) and intermediate low (IL) ranges presented in ref. 23 are shaded grey, with the mid-point (dashed line). **c**, Tropical cyclone activities (adapted from

ref. 82). Each rectangular line represents a tropical cyclone reconstruction (see ref. 82 for references for each individual reconstruction) with location grouped by North West Atlantic, red; North West Pacific, blue; South West Pacific, green; and South Indian, orange. Black represents active tropical cyclone periods and light shading less active periods. Sedimentary reconstructions of tropical cyclones exist only for the past few millennia, partly because coastlines were generally more unstable before this period due to increased rates of SLR.

Tropical-cyclone-derived beach-ridge deposits, therefore, highlight the ability of some coastlines to generally advance seaward over a period of varying tropical cyclone activity, with significant changes in the long-term behaviour of this coastal system driven not by changes in storm activity, but rather by the mid-Holocene transition from rapid rates of SLR during the Holocene transgression to stable or moderate rates of sea-level fall over the past few millennia.

Regional landscapes that were flooded during the Holocene transgression often vary in composition and geometry compared with today's coasts. Thus, the future response of these shorelines to rapid SLR will probably differ somewhat to responses during the early Holocene. However, the marked difference in form and behaviour of most of the world's low-lying sedimentary coastlines during past rapid SLR over the Holocene transgression is a clear example of the importance of sea-level variability in initiating significant changes in shoreline behaviour and, thus, should not be overlooked.

Managing future risk

By the end of this century there will probably be a higher occurrence of more intense tropical cyclones globally². However, considerable uncertainty is associated with how the smaller subset of landfalling tropical cyclones will change in the future. Efforts are ongoing to provide more robust projections of the occurrence and intensity of these events. Nonetheless, current uncertainties around the effect of future climate change on tropical cyclone activity should not distract from the two additional forces that will drive higher flood probabilities. First, increasing rates of SLR will increase extreme flooding by tropical cyclones. Second, future storm damage will be greatest not where tropical cyclone activity is the

highest, but rather where geomorphic changes along dynamic, populated shorelines greatly enhance storm impacts.

Most coastal populations are not prepared for an increase in extreme flood frequency. Coastal planners and policy makers are challenged by large uncertainties in flood projections related to changing tropical cyclone climatology, SLR and shoreline change. However, despite these uncertainties, the high likelihood of increased catastrophic coastal flooding in the future warrants preparation. Projected increases in coastal development and population will only increase damages from tropical cyclones⁵. Coastal populations need to develop adaptive strategies, which in many cases must include plans and incentives for landward or vertical retreat from the sea. Equally important is the development of proactive policies for planning and engineering in communities that must remain in these vulnerable areas, because of, for example, economic importance, national security or political boundaries. When coastal defences are necessary to protect crucial infrastructure, it is important that they are designed in a way that allows for future modification — because flooding risks will continue to increase over time as SLR accelerates through the twenty-first century (Fig. 4b). Crucial for increasing resilience to the effects of future tropical cyclones are holistic strategies that include consideration of the issues related to changes in sediment supply and subsidence induced by groundwater, oil and gas extraction. Such strategies will be particularly important along and behind barrier beaches as well as for the major deltaic systems on which many coastal megacities exist (Fig. 3).

Coastal communities in developing countries are possibly the most susceptible populations to the adverse effects of increased tropical cyclone flooding^{35,84,85}. Here, urban centres and their projected growth

are generally focused on coastal areas where existing infrastructure and current management strategies are ill equipped for extreme tropical cyclone flooding. In terms of the number of people affected, the impact of future tropical cyclone flooding will probably be focused on key population centres built on broad, low-lying sedimentary coasts⁸⁶ (Fig. 3). Essential strategies for mitigating risk at these locations include improving flood forecasts and developing emergency shelter and effective evacuation procedures⁸⁵.

Humans have adapted to environmental changes in the past. When reacting to a growing hazard, however, it is important to understand its root cause. It is possible that changes in future tropical cyclone activity could be an important component of flood risk, and management strategies will need to be updated as the science advances on this important topic. The evidence is now clear, however, that sea levels are rising and at a rate that will continue to accelerate into the next century. The era of relatively moderate SLR that most coastlines have experienced during the past few millennia is over, and shorelines are now beginning to adjust to a new boundary condition that in most cases serves to accelerate rates of shoreline retreat. The potential for future tropical cyclones to increase in their intensity has served as a prominent example of increased risk that is associated with climate change. This has placed a disproportionate emphasis on still uncertain changes to tropical cyclone characteristics at the expense of factors with a potentially larger and more certain impact, including accelerating SLR, rapidly evolving coastlines and growing coastal populations. The combined consideration of all of these elements is a much more accurate presentation of the compounding factors that society must consider to successfully adapt to future increases in tropical cyclone flooding. ■

Received 7 April; accepted 23 July 2013.

1. Peduzzi, P. *et al.* Global trends in tropical cyclone risk. *Nature Clim. Change* **2**, 289–294 (2012).
2. Knutson, T. R. *et al.* Tropical cyclones and climate change. *Nature Geosci.* **3**, 157–163 (2010).
- This article provides the most current community consensus on projections of future tropical cyclone activity.**
3. EM-DAT. *The OFDA/CRED International Disaster Database*. <http://www.emdat.be> (CRED, 2013).
4. Pielke, R. A. *et al.* Normalized hurricane damage in the United States: 1900–2005. *Nat. Hazards Rev.* **9**, 29–42 (2008).
5. Mendelsohn, R., Emanuel, K., Chonabayashi, S. & Bakkensen, L. The impact of climate change on global tropical cyclone damage. *Nature Clim. Change* **2**, 205–209 (2012).
6. Frank, W. M. & Young, G. S. The interannual variability of tropical cyclones. *Mon. Weath. Rev.* **135**, 3587–3598 (2007).
7. Weinkle, J., Maue, R. & Pielke, R. Jr. Historical global tropical cyclone landfalls. *J. Clim.* **25**, 4729–4735 (2012).
8. Gray, W. M. in *Meteorology Over the Tropical Oceans* (ed. Shaw, D. B.) 155–218 (Royal Meteorological Society, 1979).
9. Emanuel, K. A. The maximum intensity of hurricanes. *J. Atmos. Sci.* **45**, 1143–1155 (1988).
- This article presents a theoretical foundation for the direct relationship between SST and the intensity of tropical cyclones.**
10. Camargo, S. J., Emanuel, K. A. & Sobel, A. H. Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis. *J. Clim.* **20**, 4819–4834 (2007).
11. Tippett, M. K., Camargo, S. J. & Sobel, A. H. A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *J. Clim.* **24**, 2335–2357 (2011).
12. Gray, W. M. Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb Quasi-Biennial Oscillation influences. *Mon. Weath. Rev.* **112**, 1649–1668 (1984).
13. Frank, W. M. & Ritchie, E. A. Effects of vertical wind shear on the intensity and structure of numerically simulated hurricanes. *Mon. Weath. Rev.* **129**, 2249–2269 (2001).
14. Villarini, G. & Vecchi, G. A. Twenty-first-century projections of North Atlantic tropical storms from CMIP5 models. *Nature Clim. Change* **2**, 604–607 (2012).
15. Villarini, G. & Vecchi, G. A. Projected increases in North Atlantic tropical cyclone intensity from CMIP5 models. *J. Clim.* **26**, 3231–3240 (2013).
16. Kim, J.-H., Ho, C.-H., Kim, H.-S., Sui, C.-H. & Park, S. K. Systematic variation of summertime tropical cyclone activity in the western North Pacific in relation to the Madden–Julian oscillation. *J. Clim.* **21**, 1171–1191 (2008).
17. Barrett, B. S. & Leslie, L. M. Links between tropical cyclone activity and Madden–Julian Oscillation phase in the North Atlantic and northeast Pacific basins. *Mon. Weath. Rev.* **137**, 727–744 (2009).
18. Stevenson, S. Significant changes to ENSO strength and impacts in the twenty-first century: results from CMIP5. *Geophys. Res. Lett.* **39**, L17703 (2012).
19. Takahashi, C., Sato, N., Seiki, A., Yoneyama, K. & Shirooka, R. Projected future change of MJO and its extratropical teleconnection in east Asia during the northern winter simulated in IPCC AR4 models. *SOLA* **7**, 201–204 (2011).
20. Camargo, S. J., Sobel, A. H., Barnston, A. G. & Klotzbach, P. J. in *Global Perspectives on Tropical Cyclones: From Science to Mitigation*, Vol. 4 (eds Chan, J. C. L. & Kepert, J. D.) (World Scientific Publishing Company, 2010).
21. Jones, S. C. *et al.* The extratropical transition of tropical cyclones: forecast challenges, current understanding, and future directions. *Weather Forecast.* **18**, 1052–1092 (2003).
22. Kossin, J. P. & Camargo, S. J. Hurricane track variability and secular potential intensity trends. *Clim. Change* **97**, 329–337 (2009).
23. Parris, A. *et al.* *Global Sea Level Rise Scenarios for the US National Climate Assessment*. NOAA Tech Memo OAR CPO-1 (NOAA, 2012).
24. Woodworth, P. & Player, R. The permanent service for mean sea level: an update to the 21st century. *J. Coast. Res.* **19**, 287–295 (2003).
25. Menéndez, M. & Woodworth, P. L. Changes in extreme high water levels based on a quasi-global tide-gauge data set. *J. Geophys. Res.* **115**, C10011 (2010).
26. Zhang, K., Douglas, B. C. & Leatherman, S. P. Twentieth-century storm activity along the US east coast. *J. Clim.* **13**, 1748–1761 (2000).
27. Irish, J. L., Resio, D. T. & Divoky, D. Statistical properties of hurricane surge along a coast. *J. Geophys. Res.* **116**, C10007 (2011).
28. Resio, D. T. & Westerink, J. J. Modeling the physics of storm surges. *Phys. Today* **61**, 33 (2008).
29. Nicholls, R. J. & Cazenave, A. Sea-level rise and its impact on coastal zones. *Science* **328**, 1517–1520 (2010).
- This article outlines future challenges for world regions most vulnerable to future sea-level rise and subsidence.**
30. Han, M., Hou, J. & Wu, L. Potential impacts of sea-level rise on China's coastal environment and cities: a national assessment. *J. Coast. Res.* **14**, 79–95 (1995).
31. Knutson, T. R. & Tuleya, R. E. Impact of CO₂-induced warming on simulated hurricane intensity and precipitation: sensitivity to the choice of climate model and convective parameterization. *J. Clim.* **17**, 3477–3495 (2004).
32. Knutson, T. R. & Tuleya, R. E. In: *Climate Extremes and Society* (eds Diaz, H. F. & Murnane, R. J.) 120–144 (2008).
33. Emanuel, K. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **436**, 686–688 (2005).
34. Nicholls, R. J., Hoozemans, F. M. J. & Marchand, M. Increasing flood risk and wetland losses due to global sea-level rise: regional and global analyses. *Glob. Environ. Change* **9**, S69–S87 (1999).
35. Hanson, S. *et al.* A global ranking of port cities with high exposure to climate extremes. *Clim. Change* **104**, 89–111 (2011).
36. Ali, A. Climate change impacts and adaptation assessment in Bangladesh. *Clim. Res.* **12**, 109–116 (1999).
37. Church, J. A., Hunter, J. R., McInnes, K. L. & White, N. J. Sea-level rise around the Australian coastline and the changing frequency of extreme sea-level events. *Aust. Meteorol. Mag.* **55**, 253–260 (2006).
38. Irish, J. L. & Resio, D. T. A method for estimating future hurricane flood probabilities and associated uncertainty. *J. Waterw. Port Coast. Ocean Eng.* **139**, 126–134 (2013).
39. Lin, N., Emanuel, K., Oppenheimer, M. & Vanmarcke, E. Physically based assessment of hurricane surge threat under climate change. *Nature Clim. Change* **2**, 462–467 (2012).
- This study provides a rigorous evaluation for the combined influence of SLR and future tropical cyclone climate on storm surge probabilities.**
40. Smith, J. M., Cialone, M. A., Wamsley, T. V. & McAlpin, T. O. Potential impact of sea level rise on coastal surges in southeast Louisiana. *Ocean Eng.* **37**, 37–47 (2010).
- This is one of a number of important studies that quantify the nonlinear effects on surge by SLR.**
41. Rodolfo, K. S. & Siringan, F. P. Global sea-level rise is recognized, but flooding from anthropogenic land subsidence is ignored around northern Manila Bay, Philippines. *Disasters* **30**, 118–139 (2006).
42. Nicholls, R. J. Coastal megacities and climate change. *GeoJournal* **37**, 369–379 (1995).
43. Wang, J., Gao, W., Xu, S. & Yu, L. Evaluation of the combined risk of sea level rise, land subsidence, and storm surges on the coastal areas of Shanghai, China. *Clim. Change* **115**, 537–558 (2012).
44. Neumann, J. E., Emanuel, K. A., Ravela, S., Ludwig, L. C. & Verly, C. *WP 2012/81 Risks of Coastal Storm Surge and the Effect of Sea Level Rise in the Red River Delta, Vietnam* (UNU–WIDER, 2012).
45. Hoffman, R. N. *et al.* An estimate of increases in storm surge risk to property from sea level rise in the first half of the twenty-first century. *Weather Clim. Soc.* **2**, 271–293 (2010).
46. Uehara, K., Scourse, J. D., Horsburgh, K. J., Lambeck, K. & Purcell, A. P. Tidal evolution of the northwest European shelf seas from the Last Glacial Maximum to the present. *J. Geophys. Res.* **111**, C09025 (2006).
47. Hughes, T. P. *et al.* Climate change, human impacts, and the resilience of coral reefs. *Science* **301**, 929–933 (2003).
48. Hoegh-Guldberg, O. *et al.* Coral reefs under rapid climate change and ocean acidification. *Science* **318**, 1737–1742 (2007).
49. Fitzgerald, D. M., Fenster, M. S., Argow, B. A. & Buynevich, I. V. Coastal impacts due to sea-level rise. *Annu. Rev. Earth Planet. Sci.* **36**, 601–647 (2008).
- This paper reviews a century of research on shoreline change in response to changes in sea level.**

50. Goodbred, S. L. Jr, Wright, E. E. & Hine, A. C. Sea-level change and storm-surge deposition in a late Holocene Florida salt marsh. *J. Sediment. Res.* **68**, 240–252 (1998).
51. Friedrichs, C. T. & Perry, J. E. Tidal salt marsh morphodynamics: a synthesis. *J. Coast. Res.* **27**, 7–37 (2001).
52. Stumpf, R. P. The process of sedimentation on the surface of a salt marsh. *Estuar. Coast. Shelf Sci.* **17**, 495–508 (1983).
53. Cooper, M. J. P., Beevers, M. D. & Oppenheimer, M. The potential impacts of sea level rise on the coastal region of New Jersey, USA. *Clim. Change* **90**, 475–492 (2008).
54. Lacombe, P. & Carter, R. Cyclone pumping, sediment partitioning and the development of the Great Barrier Reef shelf system: a review. *Quat. Sci. Rev.* **23**, 107–135 (2004).
55. Nott, J. Tropical cyclones and the evolution of the sedimentary coast of northern Australia. *J. Coast. Res.* **22**, 49–62 (2006).
56. Cooper, J. A. G. & Pilkey, O. H. Sea-level rise and shoreline retreat: time to abandon the Bruun Rule. *Global Planet. Change* **43**, 157–171 (2004).
57. Morton, R. A., Paine, J. G. & Gibeau, J. C. Stages and durations of post-storm beach recovery, southeastern Texas coast, USA. *J. Coast. Res.* **10**, 884–908 (1994).
58. Ranasinghe, R., Duong, T. M., Uhlenbrook, S., Roelvink, D. & Stive, M. Climate-change impact assessment for inlet-interrupted coastlines. *Nature Clim. Change* **3**, 83–87 (2012).
59. Morton, R. A. & Sallenger, A. H. Jr. Morphological impacts of extreme storms on sandy beaches and barriers. *J. Coast. Res.* **19**, 560–573 (2003).
60. Wamsley, T. V., Cialone, M. A., Smith, J. M., Ebersole, B. A. & Grzegorzewski, A. S. Influence of landscape restoration and degradation on storm surge and waves in southern Louisiana. *Nat. Hazards* **51**, 207–224 (2009).
61. Fagherazzi, S., Carniello, L., D'Alpaos, L. & Defina, A. Critical bifurcation of shallow microtidal landforms in tidal flats and salt marshes. *Proc. Natl Acad. Sci. USA* **103**, 8337–8341 (2006).
62. Mariotti, G. & Fagherazzi, S. Critical width of tidal flats triggers marsh collapse in the absence of sea-level rise. *Proc. Natl Acad. Sci. USA* **110**, 5353–5356 (2013).
63. Zhang, K., Douglas, B. & Leatherman, S. Do storms cause long-term beach erosion along the US East Barrier Coast? *J. Geol.* **110**, 493–502 (2002).
- This article presents evidence for the dominance of sea-level rise and variations of sediment supply in driving long-term rates of shore-line retreat.**
64. Harmelin-Vivien, M. L. The effects of storms and cyclones on coral reefs: a review. *J. Coast. Res.* **12**, 211–231 (1994).
65. Wang, P. *et al.* Morphological and sedimentological impacts of Hurricane Ivan and immediate poststorm beach recovery along the northwestern Florida barrier-island coasts. *J. Coast. Res.* **22**, 1382–1402 (2006).
66. Done, T. J. Coral community adaptability to environmental change at the scales of regions, reefs and reef zones. *Am. Zool.* **39**, 66–79 (1999).
67. Donoghue, J. F. Sea level history of the northern Gulf of Mexico coast and sea level rise scenarios for the near future. *Clim. Change* **107**, 17–33 (2011).
68. Emery, K., Wigley, R. & Rubin, M. A submerged peat deposit off the Atlantic coast of the United States. *Limnol. Oceanogr.* **10**, R97–R102 (1965).
69. Field, M. E., Meisburger, E. P., Stanley, E. A. & Williams, S. J. Upper Quaternary peat deposits on the Atlantic inner shelf of the United States. *Geol. Soc. Am. Bull.* **90**, 618–628 (1979).
70. Pluet, J. & Pirazzoli, P. *World Atlas of Holocene Sea-Level Changes* Vol. 58 (Elsevier, 1991).
71. Stanley, D. J. & Warne, A. G. Worldwide initiation of Holocene marine deltas by deceleration of sea-level rise. *Science* **265**, 228–231 (1994).
72. Kraft, J. C. Sedimentary facies patterns and geologic history of a Holocene marine transgression. *Geol. Soc. Am. Bull.* **82**, 2131–2158 (1971).
- This article provides evidence for the landward transgression and reworking of the continental shelf by rapid rates of sea-level rise during the early Holocene.**
73. Anderson, J., Milliiken, K., Wallace, D., Rodriguez, A. & Simms, A. Coastal impact underestimated from rapid sea-level rise. *Eos* **91**, 205–206 (2010).
74. Rhodes, E. Depositional model for a chenier plain, Gulf of Carpentaria, Australia. *Sedimentology* **29**, 201–221 (1982).
75. Otvos, E. G. Coastal barriers, Gulf of Mexico: Holocene evolution and chronology. *J. Coast. Res.* **42**, 141–163 (2005).
76. Redfield, A. C. Development of a New England salt marsh. *Ecol. Monogr.* **42**, 201–237 (1972).
77. Newman, W. S. & Rusnak, G. A. Holocene submergence of the eastern shore of Virginia. *Science* **148**, 1464–1466 (1965).
78. Ellison, J. C. & Stoddart, D. R. Mangrove ecosystem collapse during predicted sea-level rise: Holocene analogues and implications. *J. Coast. Res.* **7**, 151–165 (1991).
79. Parkinson, R. W., DeLaune, R. D. & White, J. R. Holocene sea-level rise and the fate of mangrove forests within the wider Caribbean region. *J. Coast. Res.* **10**, 1077–1086 (1994).
80. Mann, M., Woodruff, J., Donnelly, J. & Zhang, Z. Atlantic hurricanes and climate over the past 1,500 years. *Nature* **460**, 880–883 (2009).
81. Woodruff, J. D., Donnelly, J. P., Emanuel, K. & Lane, P. Assessing sedimentary records of paleohurricane activity using modeled hurricane climatology. *Geochem. Geophys. Geosyst.* **9**, Q09V10 (2008).
82. Nott, J. & Forsyth, A. Punctuated global tropical cyclone activity over the past 5,000 years. *Geophys. Res. Lett.* **39**, L14703 (2012).
83. Lewis, S. E., Sloss, C. R., Murray-Wallace, C. V., Woodroffe, C. D. & Smithers, S. G. Post-glacial sea-level changes around the Australian margin: a review. *Quat. Sci. Rev.* **74**, 115–138 (2013).
84. Dasgupta, S., Laplante, B., Murray, S. & Wheeler, D. Exposure of developing countries to sea-level rise and storm surges. *Clim. Change* **106**, 567–579 (2011).
85. Webster, P. J. Meteorology: Improve weather forecasts for the developing world. *Nature* **493**, 17–19 (2013).
86. Brecht, H., Dasgupta, S., Laplante, B., Murray, S. & Wheeler, D. Sea-level rise and storm surges: High stakes for a small number of developing countries. *J. Environ. Dev.* **21**, 120–138 (2012).
87. Jarvinen, B. R., Neuman, C. & Davis, M. NOAA Tech. Memo. NWS NHC-22, A Tropical Cyclone Data Tape for the North Atlantic basin (NOAA, 1988).
88. Chu, J.-H., Sampson, C. R., Levine, A. S. & Fukada, E. *The Joint Typhoon Warning Center Tropical Cyclone Best-Tracks, 1945–2000* (Naval Research Laboratory, 2002).
89. McTaggart-Cowan, R. *et al.* Analysis of hurricane Catarina (2004). *Mon. Weath. Rev.* **134**, 3029–3053 (2006).
90. Permanent Service for Mean Sea Level. *Obtaining Tide Gauge Data*. <http://www.psmsl.org/data/obtaining/> (PSMSL, 2013).
91. United Nations. *World Urbanization Prospects, The 2011 Revision*. <http://esa.un.org/unup/> (United Nations, 2012).
92. Karim, M. F. & Mimura, N. Impacts of climate change and sea-level rise on cyclonic storm surge floods in Bangladesh. *Glob. Environ. Change* **18**, 490–500 (2008).
93. Huang, Z., Zong, Y. & Zhang, W. Coastal inundation due to sea level rise in the Pearl River Delta, China. *Nat. Hazards* **33**, 247–264 (2004).
94. NOAA. *Sea Level Trends*. <http://tidesandcurrents.noaa.gov/sltrends/> (NOAA, 2013).
95. Amante, C. & Eakins, B. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis* (DOC/NOAA/NESDIS/NGDC, 2008).
96. Fleming, K. *et al.* Refining the eustatic sea-level curve since the Last Glacial Maximum using far-and intermediate-field sites. *Earth Planet. Sci. Lett.* **163**, 327–342 (1998).
97. Milne, G. A., Long, A. J. & Bassett, S. E. Modelling Holocene relative sea-level observations from the Caribbean and South America. *Quat. Sci. Rev.* **24**, 1183–1202 (2005).
98. Peltier, W. R. On eustatic sea level history: last glacial maximum to Holocene. *Quat. Sci. Rev.* **21**, 377–396 (2002).
99. Pugh, D. *Changing Sea Levels: Effects of Tides, Weather and Climate* (Cambridge Univ. Press, 2004).
100. Scileppi, E. & Donnelly, J. P. Sedimentary evidence of hurricane strikes in western Long Island, New York. *Geochem. Geophys. Geosyst.* **8**, Q06011 (2007).

Acknowledgements We wish to thank our colleagues for the many comments and suggestions that improved this manuscript, as well as thoughtful discussions at the 2013 Joint AGU/GSA Conference on 'Coastal Processes and Environments Under Sea-Level Rise and Changing Climate: Science to Inform Management'. J.D.W. is funded through the US National Science Foundation (NSF, grant number EAR-1158780 and EAR-1148244), the Risk Prediction Initiative at the Bermuda Institute of Ocean Sciences (grant number RPI11-1-001/11-5110), and the Hudson River Foundation. S.J.C. acknowledges funding from the National Oceanic and Atmospheric Administration (NOAA, grant number NA11OAR4310093 and NA10OAR4310124) and NSF (grant number AGS-1143959 and AGS-1064081). J.L.I. received funding for this work through NOAA's National Sea Grant College Program (grant number 24036078) and the South Atlantic Landscape Conservation Cooperative (grant number 24036078). The views expressed herein do not necessarily reflect the views of any of these organizations.

Author Information Reprints and permissions information is available at www.nature.com/reprint. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/f6rg4i. Correspondence should be addressed to J.W. (woodruff@geo.umass.edu).

Tidal wetland stability in the face of human impacts and sea-level rise

Matthew L. Kirwan¹ & J. Patrick Megonigal²

Coastal populations and wetlands have been intertwined for centuries, whereby humans both influence and depend on the extensive ecosystem services that wetlands provide. Although coastal wetlands have long been considered vulnerable to sea-level rise, recent work has identified fascinating feedbacks between plant growth and geomorphology that allow wetlands to actively resist the deleterious effects of sea-level rise. Humans alter the strength of these feedbacks by changing the climate, nutrient inputs, sediment delivery and subsidence rates. Whether wetlands continue to survive sea-level rise depends largely on how human impacts interact with rapid sea-level rise, and socio-economic factors that influence transgression into adjacent uplands.

Coastal wetlands are simultaneously some of the most vulnerable and most economically important ecosystems on Earth. Marshes and mangroves protect coastal regions from storms, sequester carbon, transform nutrients and provide the organic matter and nursery grounds that support commercial fisheries¹. Although these ecosystem services are valued at about US\$10,000 per hectare¹, around 25–50% of the world's coastal tidal wetlands have been lost as a result of their direct conversion into land for agriculture and aquaculture uses^{2–4}. Tidal wetland conversion to open water through sea-level rise is expected to accelerate, with regional assessments predicting a 20–45% loss of salt marsh during the current century⁵. However, forecasts of widespread wetland loss are difficult to defend on the basis of past accelerations of sea-level rise. There are relatively few examples of marsh loss in the historical record that are directly attributable to sea-level rise because feedbacks between flooding, plant growth and elevation change tend to stabilize submerging wetlands^{6,7}. In fact, most coastal wetlands build vertically at rates similar to or that exceed the rate of historical sea-level rise^{8,9}. Regions of the world with drastic wetland deterioration occur mainly in areas in which humans have accelerated subsidence rates and/or decreased sediment delivery rates to the coast (for example, coastal Louisiana, the Venice Lagoon and Chesapeake Bay). Nevertheless, past response to sea-level rise is an imperfect model for future response because the climate, water quality and sediment delivery rates continue to change with human activity. In this Review, we will discuss the processes that influence how tidal wetlands adapt to sea-level rise, and highlight how changing climate and socio-economic conditions may alter our emerging understanding of riveting feedbacks between ecology and geomorphology. We focus mainly on tidal marsh ecosystems for which the ecogeomorphic feedbacks are better understood, but also note instances in which data or general principles apply to mangroves. We argue that human impacts other than those that cause sea-level rise have dominated wetlands in the past, but that interactions between rapid sea-level rise and human impacts will drive wetland stability in the future. Whether these ecosystems continue to survive ever faster rates of sea-level rise depends principally on sediment availability, biotic responses to environmental change, the opportunity for wetlands to migrate inland, and environmental attitudes that influence land use, all of which are heavily determined by human socio-economic systems.

Biophysical feedbacks stabilize wetlands

Expansive tidal wetlands consisting of marshes and mangroves, and the channel networks that dissect them occupy about 20 million hectares worldwide³, and have been a prominent component of coastal and estuarine landscapes for at least 4,000 years¹⁰. Over this period, the sea level has risen in most regions of the world by more than 2 metres^{11,12}. However, observations of widespread wetland drowning are infrequent because of the fascinating interactions between plants and soil that allow wetlands to actively engineer their position within the intertidal zone in ways that enhance ecosystem persistence^{7,13–15}.

Vertical changes in wetland elevation

At the most basic level, a marsh or mangrove must build soil elevation at a rate faster than or equal to the rate of sea-level rise to survive in place¹⁶. Elevation gain occurs through biological and physical feedbacks that couple the rate of sea-level rise to the rate of vertical accretion (the increase in soil surface elevation) (Fig. 1). In their role as ecosystem engineers, plants set up distinct feedback loops above and below ground. Above ground, mineral sediment settles out of the water column and onto coastal wetland soils during periods of tidal flooding, so that deposition rates are highest in low elevation marshes that are inundated for long periods of time, and lowest in high elevation marshes that are more rarely flooded^{17,18} (Fig. 2a). Plant shoots influence mineral sediment deposition by slowing water velocities⁷, and add organic matter to the soil surface (Fig. 1). Below ground, the balance of plant root growth and decay directly adds organic matter to the soil profile, raising elevation by sub-surface expansion¹⁹.

Coastal wetlands are among the most productive ecosystems on Earth, and recent work suggests that vegetation tends to stabilize their relative elevation and seaward extent through feedbacks that vary with the depth and duration of flooding. For example, growth of the grass *Spartina alterniflora* is positively correlated with interannual variations in sea level, such that productivity peaks at intermediate elevations within the intertidal zone, and declines at higher or lower elevations²⁰ (Fig. 2a). Although the response of mangrove productivity to interannual sea-level variation is unknown, other marsh species show similar — but species-specific — patterns^{21,22}. Faster rates of above-ground plant growth promote greater standing biomass, which in turn slows water velocities on the marsh platform²³, lowers wave height²⁴, reduces erosion and enhances mineral sediment deposition²⁵. Collectively, these feedbacks allow tidal

¹Virginia Institute of Marine Science, PO Box 1346, 1375 Greate Road, Gloucester Point, Virginia 23062, USA. ²Smithsonian Environmental Research Center, 647 Contees Wharf Road, Edgewater, Maryland 21037, USA.

marshes to survive accelerating rates of sea-level rise^{6,20}. Similar feedbacks between flooding, plant growth and sub-surface expansion operate in the root zone, generating highly organic soils that persist for thousands of years^{19,21,26} (Box 1). Together, these eco-geomorphic interactions suggest that more extensive flooding associated with sea-level rise should be accompanied by enhanced accretion. Indeed, vertical accretion rates approximately tripled in several marshes surrounding Long Island, New York, in response to twentieth century sea-level acceleration²⁷.

Spatial landscape-scale feedbacks

Landscape-scale geomorphic processes are also important in determining the stability of coastal wetlands. In regions where subsidence is limited and vertical drowning is relatively uncommon^{8,9}, the size of today's wetlands largely reflects the difference between the rate of lateral erosion at the seaward margin²⁸, and the rate of wetland creation (that is, migration) at the landward margin (Fig. 1). Erosion rates tend to increase with sea-level rise in shallow intertidal environments because increases in water depth reduce the amount of dissipation that occurs as incoming waves move across tidal flats²⁹. Preliminary work suggests that rates of wetland expansion into adjacent forests may accelerate with future sea-level rise^{30,31}. Therefore, coupling ecological models of the marsh–forest margin with geomorphic models of retreat at the seaward edge is an important direction for future research.

Sediment dynamics in submerging coastal landscapes can aid vertical accretion in tidal wetlands by delivering sediment from eroding portions of the landscape and depositing it in other portions. For example, rapid erosion of subtidal flats provides sediment to adjacent wetlands on the Yangtze River delta, China, allowing marshes to maintain their aerial extent³². Similarly, expansion of channel networks in response to accelerated sea-level rise may deliver more sediment to portions of the platform that were previously sediment deficient^{33,34}. Together, these types of ecogeomorphic feedbacks probably explain the persistence of wetlands within the intertidal zone over thousands of years in the stratigraphic record¹², and observations of accretion rates that are highest in regions with historically high rates of sea-level rise¹³.

Threshold rates of sea-level rise

Despite robust ecogeomorphic feedbacks that stabilize tidal wetlands, observations of wetland deterioration in places such as the Mississippi River Delta indicate that there are limits to the feedbacks that preserve wetlands within the intertidal zone. An emerging idea is that marshes survive increasing rates of sea-level rise by becoming lower in the tidal zone, which allows them to build elevations at progressively faster rates until they become so flooded that vegetation dies off, and stabilizing ecogeomorphic feedbacks are lost^{6,20}. However, the rate of sea-level rise beyond which marshes tend to drown is highly site specific and heavily influenced by human impact, ranging from a few millimetres to several centimetres per year⁶ (Fig. 2).

Wetland deterioration

Large areas of marsh are being converted to open water in the Gulf of Mexico, Venice Lagoon and along tributaries of the Chesapeake Bay^{16,35,36}. In these regions, which are characterized by low elevations and/or fast rates of relative sea-level rise, increases in the duration of tidal inundation no longer stimulate plant productivity. Rather, progressive inundation reduces organic matter contributions from plants and accelerates erosion, causing a feedback that accelerates the deterioration of coastal wetlands^{20,37,38} (Fig. 2a). A variety of numerical models suggest that the transition from a stable to unstable marsh is mainly regulated by the tidal range of an estuary (which sets the elevation range over which plants can grow) and the amount of sediment available for marsh accretion⁶ (Fig. 2b). Each of these rapidly deteriorating systems is located in an estuary with small tidal ranges and sediment inputs.

A fundamental goal of tidal wetland research is to forecast the conditions under which tidal wetlands undergo a state change to open water or mud flat, and to relate this back to threshold rates of sea-level rise that can be measured and monitored. The geological record offers some insight. Submerged salt marshes are often preserved as layers of organic rich peat in the stratigraphy of bays, estuaries and the offshore continental shelf^{10,39–41}. Although more work is needed to connect the collapse of these palaeo-marshes with historical rates of sea-level rise, peat from modern

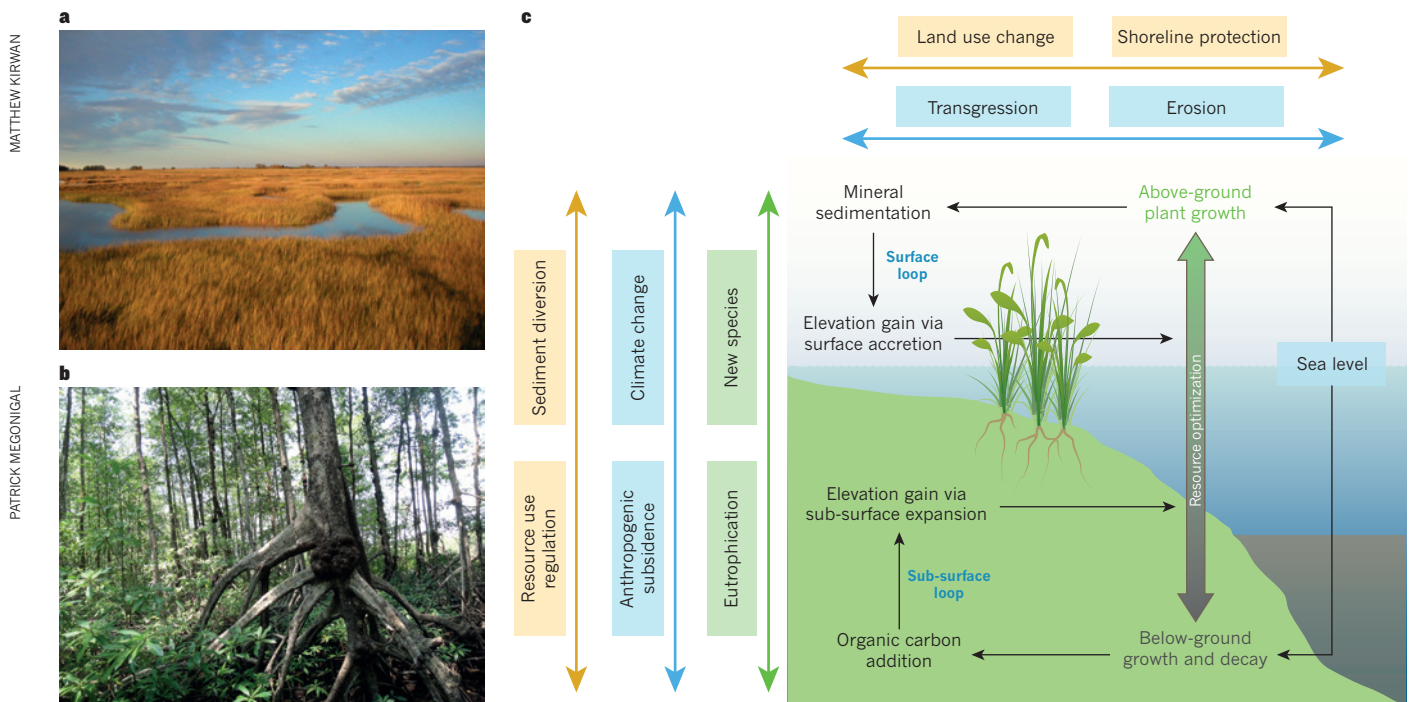


Figure 1 | Wetland feedbacks. Feedbacks in marshes (top left) and mangroves (bottom left) operate horizontally and vertically at different scales and with distinct sets of processes to influence the wetland stability. Feedbacks on vertical elevation change operate through natural processes

above and below ground. These natural processes can be perturbed by local factors (green) such as eutrophication and new species; large-scale climatic and geomorphic processes (blue); and political, social and economic factors (orange), which affect the other processes.

marshes suggests that marshes form and persist when relative sea level rises at a rate of less than a couple of millimetres per year¹², but that existing marshes survive much faster rates. Most (>90%) basal peats from salt marshes along the US Atlantic Coast are less than 6,000 years old, implying that most modern marshes formed during a time when relative sea-level rise rates were slowing from 1–4 mm yr⁻¹ to 0.5–2 mm yr⁻¹ (ref. 12). This suggests that marshes mainly establish when rates of relative sea-level rise are quite low. However, rates of sea-level rise at marsh inception are a minimal estimate of threshold rates for survival because biophysical feedbacks (Fig. 1) allow established marshes to survive conditions in which they cannot form^{42,43}. For example, mid-Holocene marshes that responded to rapid sea-level rise 8,200 ybp survived rates of about 7 mm yr⁻¹ in Louisiana⁴⁴, and drowned in Chesapeake Bay only when rates exceeded 12 mm yr⁻¹ (ref. 45).

Historical persistence

The response of salt marshes to sea-level rise can also be viewed in the context of more recent sea-level acceleration. Tide gauges and stratigraphic evidence indicate that relative sea-level rise rates were less than 1 mm yr⁻¹ for most of the past 2,000 years, and began accelerating towards modern rates (about 2–3 mm yr⁻¹) around the end of the nineteenth century¹¹. Perhaps in response, more flood-tolerant vegetation such as *Spartina alterniflora* invaded New England marshes, which had historically been dominated by flood-intolerant vegetation such as *Spartina patens*, at roughly the same time sea-level rise began to accelerate⁴⁶. Although these are local observations, numerical models indicate that historical sea-level-rise acceleration would have led to a modest (around 5–15 cm) deepening of marsh surfaces relative to sea level⁴⁷. Nevertheless, most models predict threshold rates of sea-level rise (5–50 mm yr⁻¹) that are much faster than what has occurred in the recent past⁶ (Fig. 2b).

Measurements of vertical accretion rates in tidal wetlands around the world are consistent with models that predict relatively fast threshold rates of sea-level rise. Although some tidal wetlands are flooded for longer durations, as evidenced by changes in vegetation type, there seems to be no evidence of widespread wetland loss that is directly related to sea-level rise^{8,9}. These data emphasize that threshold rates of sea-level rise have rarely been crossed in recent decades. However, it remains unclear how anthropogenic impacts will shift thresholds.

Human interference with ecosystem feedbacks

Historical observations yield clues as to the maximum rate of sea-level rise that tidal wetlands can tolerate, but are ultimately limited by substantial differences between past, present and future environmental conditions. Compared with the last period of rapid sea-level rise 8,200 ybp⁴⁵, the present and future are characterized by higher atmospheric carbon dioxide concentration, plant-available nitrogen, temperature and introductions of new plant and animal species, all of which influence the major natural feedback processes that stabilize tidal wetland ecosystems (Fig. 1).

Deterioration of tidal wetlands often begins with plant stress, and the disruption of the stabilizing feedbacks that plants provide. For example, plant mortality associated with the BP Deepwater Horizon oil spill triggered order-of-magnitude increases in marsh edge erosion rates⁴⁸, historically stable channel networks became strongly erosive when crabs disturbed plants and substrate⁴⁹, herbivory caused an accreting marsh on an actively building delta to become strongly erosive³⁴, and tree mortality wrought by Hurricane Mitch caused mangrove peat collapse⁵⁰. Even temporary, climatically driven episodes of vegetation die-off^{51,52} sometimes lead to geomorphic change, including rapid subsidence, platform erosion and diminished deposition rates^{23,53}. Thus, factors that influence the growth rate of plants (for example, climate and nutrients) are likely to influence the ability of a marsh to survive sea-level rise.

Climate change and eutrophication

The effect of any given perturbation on tidal wetland stability depends a great deal on the extent to which it affects above ground compared with below-ground feedbacks (Fig. 1). For example, elevated CO₂ increases

the photosynthetic efficiency of above-ground (C₃) plant tissues, plant demand for root-acquired soil nutrients and root growth^{54,55}. Plants with C₄ photosynthetic pathways show little response to elevated atmospheric CO₂ because their photosynthetic apparatus naturally concentrates CO₂ at the site of the primary CO₂-fixing enzyme²⁶. Although elevated CO₂ may also accelerate the decay of soil organic matter, the net effect is to increase soil mass, subsurface expansion and elevation gain (Fig. 1), all of which can occur without an increase in mineral sediment deposition. Thus, elevated CO₂ probably either enhances wetland stability through increased root production (C₃-dominated wetlands such as mangroves, brackish marshes and tidal freshwater marshes) or has no effect on stability (C₄-dominated systems such as *Spartina* salt marshes).

Latitudinal gradients suggest that warming will increase tidal wetland productivity⁵⁶ and decomposition^{57,58}, with the net effect that carbon storage and vertical accretion will be enhanced — at least initially⁵⁸. The few experimental manipulations of temperature in tidal marshes confirm this pattern^{59,60}, but suggest that long-term temperature responses will be more complex owing to species replacement⁶⁰ and interactions with rates of sea-level rise⁵⁸. The effects of warming on

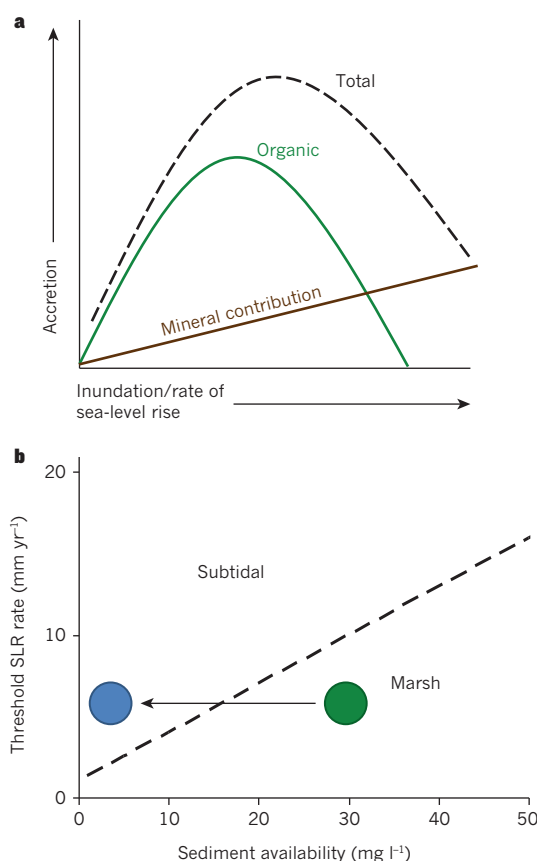


Figure 2 | Conceptual links between sea-level rise and marsh accretion. a, The hypothetical contribution of organic and mineral matter to accretion as a function of inundation in a sediment-deficient marsh. Organic matter dominates total accretion for infrequently flooded marshes that are typical of high-elevation marshes and/or periods of slow sea-level rise (left). However, the same marsh becomes progressively more mineral rich as inundation duration and rate of sea-level rise increase (right). Therefore, the threshold rate of sea-level rise tends to be a function of sediment availability. **b,** Threshold rates of sea-level rise (SLR) beyond which marshes cannot survive as a function of suspended sediment concentration in an estuary. Dashed line represents threshold rates from the 1-m tidal range case from ref. 6. Under moderately rapid sea-level rise (5 mm yr⁻¹), a marsh that is stable under historical sediment loads (green circle) submerges if sediment loads are reduced (blue circle). This suggests that land use change and dam construction may cause marshes to become less stable in the future, even if sea-level rise rates remain constant.

BOX 1

Organic contributions to elevation

Soil elevation is the result of complex interactions between the three components of soil volume: mineral matter, organic matter, and water- or gas-filled pore space. Soil accretion is sensitive to both mineral and organic deposition (Figs 1 and 2), but the ephemeral nature of organic matter makes it particularly sensitive to disturbance. Depending on the geomorphic setting, organic matter accounts for between 1 and 80% of the dry mass of tidal wetland soils, commonly forming peat soils (histosols)⁹⁴. Organic matter particles occupy about twice the volume of mineral particles on a mass-normal basis (about $0.8 \text{ cm}^3 \text{ g}^{-1}$ compared with $0.4 \text{ cm}^3 \text{ g}^{-1}$), and soil organic matter contributes 2–5 times more to bulk soil volume than an equal mass of minerals^{94,95}. Accretion rates often have a stronger correlation with organic matter accrual than mineral accrual in North American tidal marshes^{94,95}, although the reverse is sometimes true and the relationship is site and region dependent. Organic matter accrual is the main process by which tidal wetlands become perched high in the tidal frame, which reduces their vulnerability to rapid sea-level rise or decreased plant productivity.

Organic matter derived from roots, shoots and allochthonous inputs accumulate in wetland soils because a large fraction is recalcitrant to decay in the absence of oxygen, the overwhelming agent of preservation in wetland soils. The molecular composition of plant tissue is an important secondary factor, but many mechanisms of organic matter preservation in upland soils⁹⁶ are unimportant in wetlands. For

example, physical protection by mineral armouring is largely absent in organic soils and of little consequence in tidal mineral soils, which lack aggregates owing to limited fungal activity and wet-dry cycles.

There are limits to the suggestion that slow decay in wetlands is explained by the low free-energy yield of anaerobic respiration. For example, accumulation of phenolic compounds in peat-land soils can directly inhibit microbial biodegradation⁹⁷. Most effort has been devoted to the terminal steps of anaerobic decomposition, rather than the fermentation processes that precede it⁹⁸. We know little about the factors that regulate fermentative bacteria, enzyme activity, substrate feedbacks and microbial community interactions — all of which affect organic matter volume.

The delivery of salts and sulphates to brackish and freshwater coastal wetlands through sea-level rise may destabilize soil organic matter pools. Organic accretion rates tend to be highest in freshwater tidal wetlands⁹⁹, and studies report accelerated decomposition rates with saltwater intrusion¹⁰⁰, but these results are equivocal and we lack the mechanistic insight to explain such responses. Finally, most studies of decomposition focus on the decay of relatively labile, leaf and root litter over timescales of less than 3 years. The fraction of net primary production that is preserved after a decade or more is much more crucial for the accumulation of soil carbon and the maintenance of wetland elevation³⁰.

mangrove productivity are far less certain because even a relatively small rise in local temperatures (less than 1.3°C) will expose these systems to year-round temperatures well outside (more than 2 standard deviations) current variability⁶¹.

Coastal eutrophication might be expected to enhance elevation gain owing to higher rates of plant growth, but nutrient enrichment experiments show the full spectrum of elevation responses from gain to loss^{19,55,62,63}. In a single Caribbean mangrove swamp, nitrogen addition decreased or reversed elevation gain at fringe and interior sites, but had no effect on sites transitional between the two; likewise, adding phosphorus stimulated elevation gain in areas other than the fringe, at which it suppressed elevation gain⁶². In this low-sediment environment, these seemingly enigmatic responses were driven solely by below-ground processes, and mainly by differences in fine-root growth⁶², which increased, decreased or remained unchanged depending on the initial state of nitrogen and phosphorus limitation. Similar observations were reported for a peat-forming tidal marsh⁵⁵. This is in contrast with sediment-rich systems in which any increase in plant growth — root or shoot — is likely to enhance elevation gain because biomass enhances mineral sediment deposition¹⁹ (Fig. 1). Nutrient-induced elevation loss may be caused by a shift in plant growth from nutrient-acquiring roots to light-harvesting shoots, competitive replacement of a high-biomass species by a low-biomass species⁵⁴, or enhanced organic matter decay rates⁶⁴. Of these, decay responses to nutrient enrichment are the most poorly understood because studies often fail to distinguish between root respiration and soil organic-matter respiration in field studies; artificially interrupt interactions between microbial and root processes by separating the two in laboratory incubations; or focus on short-term litter decay, which has little relevance to organic-matter preservation (Box 1). Reconciling the direction of eutrophication effects on elevation will require an understanding of the processes that operate over long timescales (decades) and large areas (square kilometres). It has been suggested that eutrophication reduces soil strength in wetlands^{64,65}, but the effect of such change may take decades or a major storm event to become apparent^{64,66}. This topic is controversial⁶³ and ripe for new experimental approaches.

Experimental design limits our ability to forecast tidal wetland response

to change. The limited duration and spatial scale of most designs does not capture the tendency of ecosystems to resist perturbation until they reach a crucial threshold, after which they undergo a rapid change in state⁶⁴. The simplicity of factorial designs can be at the expense of defining response curves that are more useful for modelling. Experimental designs that support modelling are important because models can identify hysteresis or specific sets of initial conditions that influence vulnerability. For example, warming can inhibit accretion when initial rates of sea-level rise and primary production are low, or stimulate accretion when the rate of sea-level rise is initially high⁵⁸. A challenge for tidal wetland research is to define the suite of initial conditions, and interactive variables that generate complex patterns of tidal wetland stability. One such factor is plant species composition.

Vegetation shifts

The consequences of gaining or losing plant species are often more drastic than changes in the growth or physiology of existing plant species. New species influence tidal wetland stability by adding or subtracting new physiological and morphological traits that contribute to ecogeomorphic feedbacks. Low-salinity marshes of the Mississippi River delta sustained more damage from Hurricane Katrina and Hurricane Rita than high-salinity marshes because they are dominated by species with relatively shallow root profiles and consequently lower resistance to surging water and waves⁶⁶. Genotypes of the grass *Phragmites australis* introduced to North America from Europe are likely to stabilize tidal wetlands because of traits that support higher below-ground productivity than the vegetation they are replacing^{67,68}. The subsidy in soil-elevation gain provided to C_3 -dominated wetlands by elevated CO_2 can be diminished when other factors, such as eutrophication, favour C_4 species⁵⁴. As these examples show, forecasting marsh vulnerability to sea-level rise requires attention to key functional attributes of tidal wetland species such as root depth distributions and responses to perturbation.

Subsidence and sediment delivery

Humans also indirectly threaten the survival of coastal wetlands by altering subsidence rates and restricting sediment delivery (Fig. 1).

Groundwater withdrawal and artificial drainage of wetland soils contribute to rapid subsidence such that 8 of the world's 20 largest coastal cities now experience relative sea-level rise rates that greatly exceed any likely climate-driven projection⁶⁹, and most of the world's major river deltas are sinking much faster than the historical rate of sea-level rise⁷⁰. Although subsidence from isostatic flexure and the compaction of young unconsolidated sediment has a sizable natural component, subsidence caused by artificial drainage and groundwater extraction near metropolitan areas such as New Orleans, Louisiana, and Venice, Italy, can be up to an order of magnitude faster⁷¹. Temporal variations in recent subsidence rates also correlate with estimates of hydrocarbon extraction⁷². Spatial patterns of wetland loss in coastal Louisiana correlate with the density of canals built by oil and gas companies⁷³, and temporal patterns of wetland loss correlate with variation in subsidence rates⁷².

Dams and reservoirs now prevent about 20% of the global sediment load from reaching the coast⁷⁴. Because mineral sediment availability is a primary driver of wetland building, changes in sediment delivery rates have large impacts on marsh sustainability^{43,75}. An ensemble of numerical models predicts that threshold rates of sea-level rise respond linearly to changes in suspended sediment concentration, where marshes in sediment-rich estuaries survive rates of sea-level rise much greater than projected climate-driven scenarios⁶. Indeed, regions of the world with rapid wetland conversion to open water (for example, the Gulf of Mexico, Venice Lagoon and along tributaries of the Chesapeake Bay) are all located in sediment-deficient areas^{16,35,36}. Dam construction, reforestation and agricultural sediment-control practices continue to lower sediment yields to the coast⁷⁴, so these observations suggest that historically stable coastal wetlands may become increasingly prone to collapse in the future, even if sea-level rise rates were to remain steady⁶ (Fig. 2b).

Marshes on the Yangtze River delta, for example, have expanded seaward since the seventh century, surviving subsidence-generated sea-level rise rates of more than 50 mm yr⁻¹. After sediment restriction associated with the construction of more than 50,000 dams on Yangtze River tributaries, marshes in several areas are now eroding landward, and overall rates of marsh expansion have declined to near zero^{31,76}.

Direct human modification of wetlands

Direct human modification, rather than sea-level rise, is by far the major cause of historical and contemporary coastal wetland loss. Although more robust estimates are needed, conversion of wetlands into other land uses claimed about 25–50% of the world's coastal wetlands during the twentieth century alone^{2–4}. Wetland habitat conversion is an ongoing phenomenon despite several decades of investment in research, policy, education, laws and treaties aimed at understanding and conserving these resource-rich ecosystems. The history of coastal wetland degradation tracks human population growth, industrialization and development, with marginally sustainable use of coastal resources giving way to rapid decline 150–300 ybp⁷⁷. Tidal marshes were among the earliest coastal wetlands to be modified on a large scale⁷⁸ because they dominate the temperate zone where industrialization began. Intentional conversion of tidal marshes has slowed in developed countries with the adoption of laws and conservation efforts, leaving unintentional conversion to open water as the major cause of loss⁷⁹. However, developing countries are at present converting coastal wetlands to other land uses at high rates⁸⁰, substituting agriculture, aquaculture and tourism for the natural capital and ecosystem services these systems provide. For example, between 1975 and 2005, countries in the tsunami-affected



Figure 3 | Human disturbance of tidal wetland ecosystems. **a**, Fisherman's and Manatee's Cays, Belize, where mangroves were cut and filled with substrate dredged from nearby patch reefs to create white beaches. **b**, Aerial image of mangrove swamps that have been converted into shrimp ponds. **c**, Tidal marsh prevented from migrating landward by a sea wall. **d**, Subsidence of a tidal freshwater peat land in California after being dyked (embankment created), drained and farmed.

region of Asia converted 12% of their mangrove forests to agriculture and aquaculture⁸¹, despite some evidence that these systems provide protection against tsunamis and storm surge^{82–84}. A challenge is to fully quantify the socio-economic and ecological costs of wetland conversion and bio-engineering activities, and incorporate these costs in policy, planning and restoration activities^{82,85}.

Socio-economic factors

Economic incentives to expand arable land, harvest resources and protect infrastructure investments have long motivated humans to actively alter the land–sea margin⁷⁸ (Fig. 3). Such activities have generally served to degrade tidal wetlands, and to do so at an increasingly global scale that is certain to intensify with ongoing global population growth and economic development⁷⁷. The future vulnerability of tidal wetlands to degradation and loss will be a function of interacting natural and socio-economic phenomena⁸⁶ that must be reconciled through informed decision making. For example, it may be possible to simultaneously accommodate limited conversion of mangrove to shrimp ponds and maintain certain ecosystem services such as wave attenuation that scale non-linearly with wetland size⁸⁵. Thus, it is no longer sufficient to focus separately on the natural processes that sustain coastal systems, the economic incentives for human activities that disrupt these processes, and the social dimensions of human behaviour.

During the millennial period in which people's interactions with the sea have been most intense, sea-level rise rates have remained low. Only now are we beginning to learn how to respond to accelerating sea-level rise. Historical strategies for protecting coastal property have favoured use of vertical, often hardened structures such as dykes, sea walls, revetments and bulkheads^{87,88} (Fig. 3). Because intertidal wetlands lie between these structures and the sea, such measures contribute to wetland loss through 'shoreline squeeze', in which erosion removes the wetland area at the margin and structures prevent the addition of area by migration onto adjacent uplands⁸⁷. Because rates of marsh-edge erosion increase with rates of sea-level rise²⁸, the impacts of these barriers will accelerate with climate change, and the effect of coastal defence on the trajectory of coastal wetland area is potentially large. In the absence of anthropogenic barriers, a 1 m rise in sea level would create around 11,000 km² of new intertidal area in the conterminous United States alone³¹. This is a significant percentage of the existing US intertidal zone (about 16,000 km²)³¹, suggesting that sea-level-induced losses of existing wetlands may be offset by transgression if anthropogenic barriers are minimal. However, alternatives to flood defence structures that allow wetland migration require the cooperation of stakeholders on adjacent uplands, and creating these alternatives will become more difficult as the coast is developed.

The non-market value of ecosystem services is being used to promote the conservation, restoration and creation of coastal wetlands, and to protect adjacent uplands for wetland transgression. For example, the 1990 US Coastal Wetlands Planning, Protection and Restoration Act (Public Law 101-646) invests \$30–80 million annually in coastal restoration. An emerging strategy is to market the substantial capacity of coastal wetlands to store and retain carbon^{3,89}. Mangroves, salt marshes and sea grasses — blue carbon ecosystems — are global carbon hot spots where area-based carbon pools and fluxes far exceed those of other terrestrial and aquatic ecosystems⁴. Because the highest wetland loss rates and area-based carbon pools converge in mangroves, the highest potential for generating carbon credits is in developing countries where financial resources for climate mitigation are most limited. Forecasts of global wetland loss owing to sea-level rise alone are small when compared with forecasts of loss owing to the combined effects of sea-level rise and human activities related to adaptation⁸⁶. Therefore, the fate of wetlands in the twenty-first century fundamentally depends on socio-economic conditions, policy decisions and perceptions about the value of coastal wetlands^{4,90}.

Priorities for future research

For more than 30 years, point-based comparisons between rates of sea-level rise and elevation change have dominated wetland vulnerability research. However, many of the most fundamental questions pertaining to coastal wetland stability and value are inherently spatial in nature. Will wetlands transgress landward at a rate that exceeds seaward displacement? Could sea-level rise actually cause wetlands to expand? What factors explain spatial and geographical variations in tidal wetland vulnerability? To answer these questions will require integrating studies of wetland processes in the vertical dimension with research on the factors that control the lateral position of wetland boundaries. This research will require accessible sources of high-resolution digital elevation models, and data layers on the prevalence of important landscape features such as anthropogenic barriers and population density. It will also require more process-level research on the factors that control edge erosion²⁸, rates of forest-to-marsh conversion⁹¹ and land use change⁸⁸. For example, in the absence of anthropogenic barriers in the conterminous United States, preliminary work suggests that even complete drowning of existing wetlands may result in only a 22% decrease in potential wetland area because significant upland area could be available for wetland migration³¹. Thus, a systematic evaluation of the amount of land where humans restrict marsh transgression, or are likely to do so in the future, represents a simple and crucial step towards understanding whether the world's wetlands will expand or contract with sea-level rise.

In coastal regions, where the world's population continues to converge, two-way couplings between society and ecosystems are particularly captivating. Humans now have an impact on every major process influencing wetland stability (Fig. 1). Upstream land use change and dam construction alter sediment delivery rates to the coast, fluid withdrawal accelerates relative sea-level rise, eutrophication affects plant growth and decay of organic matter, and climate affects every biogeochemical process. But humans are themselves influenced by the enormous ecosystem services wetlands provide, including coastal protection from storms and rising water^{83–85}. These human impacts interact with each other, and with sea-level rise. Because of these new interactions, threshold rates of sea-level rise for marsh submergence predicted by numerical models and observed in the stratigraphic record will probably be poor indicators of future wetland vulnerability. Incorporating the indirect effects of humans on climate, sediment availability and nutrient loads into biophysical models of coastal wetland evolution is an important challenge. Indeed, preliminary work indicates that even the direction of change they induce may be site specific (for example, eutrophication). Thus, more process-level research is needed before quantitative assessments of global wetland vulnerability can hope to account for the indirect effects of human modification. For example, we have identified the processes that regulate organic matter accumulation in tidal wetland soils as one area in which more research is needed (Box 1). Large-scale manipulative experiments that push the limits of wetland survival and incorporate human actions also seem especially relevant, because most natural wetlands have adapted to historic sea-level rise alone.

Coastal population growth and accelerating rates of sea-level rise will intensify the tight interactions between society and coastal wetlands. The effect of decisions that determine how governments and landowners conserve wetlands and defend uplands from rising seas may dwarf the effect of sea-level rise alone⁸⁶. Thus, new socio-economic research examining perceptions of wetland value is needed to fully understand coastal sustainability⁹⁰. Here again, integrating direct (for example, barriers and land conversion) and indirect (for example, climate, sediment supply and nutrients) human impacts into numerical models of wetland vulnerability remains challenging. Large-scale coastal vulnerability models largely ignore the biophysical feedbacks that are known to aid marsh persistence, whereas process-oriented models are highly site specific and do not include human components⁷. The disconnect between these modelling approaches must be bridged to predict how the size and global distribution of wetlands will change in response to

climate and future human activity.

The historical loss of coastal wetlands has been dominated by the direct conversion of wetlands to agriculture and aquaculture, rather than by climate change. However, recent disasters such as Hurricane Katrina and Hurricane Sandy, the Indian Ocean tsunami and the Deepwater Horizon oil spill have renewed public interest in wetland restoration as a mechanism to provide economically valuable coastal protection⁸². Towards these efforts, our Review provides two insights. First, biophysical feedbacks allow coastal wetlands to survive conditions under which they cannot develop^{42,43}. Such a hysteresis is challenging to overcome in efforts to restore severely degraded landscapes⁹², but may bode well for the longevity of creation and restoration activities. Relying on plants to modify their environment and build wetland elevations is an intriguing strategy that should be pursued in future research. However, in some cases the biophysical environment in which these systems formed may no longer support restoration to their recent condition. For example, there is no longer enough sediment delivered to the Mississippi River Delta to fully restore the landscape to an elevation at which plants can grow and initiate these feedbacks⁹³, forcing value-laden decisions about which portions of the landscape to restore. Second, historical adaptation to sea-level rise indicates that the loss of wetlands is not an inevitable outcome of climate change. Although very rapid rates of sea-level rise may drown some marshes regardless of indirect human impacts, numerical models predict that many wetlands will survive in places in which dams and embankments do not restrict sediment transport⁶ (Fig. 2b). Preliminary topographic analyses suggest that wetland migration could largely offset even a complete loss of existing coastal wetlands in the absence of anthropogenic barriers³¹. Thus, we propose that the fate of coastal wetlands is perhaps more intrinsically linked to the complex economic and sociological decisions aimed at protecting coastal infrastructure from the impacts of climate change, than the rates and magnitude of the change itself. ■

Received 21 December 2012; accepted 11 July 2013.

- Barbier, E. B. *et al.* The value of estuarine and coastal ecosystem services. *Ecol. Monogr.* **81**, 169–193 (2011).
- Huang, Y. *et al.* Marshland conversion to cropland in northeast China from 1950 to 2000 reduced the greenhouse effect. *Glob. Change Biol.* **16**, 680–695 (2010).
- Pendleton, L. *et al.* Estimating global “blue carbon” emissions from conversion and degradation of vegetated coastal ecosystems. *PLoS ONE* **7**, e43542 (2012). **This article estimates that half of global wetlands have been lost due to direct human conversion.**
- McLeod, E. *et al.* A blueprint for blue carbon: towards an improved understanding of the role of vegetated coastal habitats in sequestering CO₂. *Front. Ecol. Environ.* **9**, 552–560 (2011).
- Craft, C. *et al.* Forecasting the effects of accelerated sea-level rise on tidal marsh ecosystem services. *Front. Ecol. Environ.* **7**, 73–78 (2009).
- Kirwan, M. L. *et al.* Limits on the adaptability of coastal marshes to rising sea level. *Geophys. Res. Lett.* **37**, L23401 (2010). **This article demonstrates that the maximum rate of sea-level rise a marsh can survive is a linear function of sediment supply and tidal range.**
- Fagherazzi, S. *et al.* Numerical models of salt marsh evolution: ecological, geomorphic, and climatic factors. *Rev. Geophys.* **50**, RG1002 (2012).
- French, J. Tidal marsh sedimentation and resilience to environmental change: exploratory modeling of tidal, sea-level, and sediment supply forcing in predominantly allochthonous systems. *Mar. Geol.* **235**, 119–136 (2006).
- Cahoon, D. R. *et al.* in *Wetlands and Natural Resource Management: Ecological Studies*, Vol. 190 (eds Verhoeven, J. T. A., Beltman, B., Bobbink, R. & Whigham, D. F.) 271–292 (Springer, 2006). **This provides a summary of elevation trends and the factors that control them from marshes around the world.**
- Rampino, M. R. & Sanders, J. E. Episodic growth of Holocene tidal marshes in the northeastern United States: a possible indicator of eustatic sea-level fluctuations. *Geology* **9**, 63–67 (1981).
- Kemp, A. C. *et al.* Climate related sea-level variations over the past two millennia. *Proc. Natl Acad. Sci. USA* **108**, 11017–11022 (2011).
- Engelhart, S. E. & Horton, B. P. Holocene sea level database for the Atlantic coast of the United States. *Quat. Sci. Rev.* **54**, 12–25 (2012).
- Friedrichs, C. T. & Perry, J. E. Tidal salt marsh morphodynamics. *J. Coast. Res.* **27**, 6–36 (2011).
- Larsen, L. G. & Harvey, J. W. How vegetation and sediment transport feedbacks drive landscape change in the Everglades and wetlands worldwide. *Am. Nat.* **176**, E66–E79 (2010).
- Marani, M., Da Lio, C. & D’Alpaos, A. Vegetation engineers marsh morphology through multiple competing stable states. *Proc. Natl Acad. Sci. USA* **110**, 3259–3263 (2013).
- Reed, D. J. The response of coastal marshes to sea-level rise: survival or submergence? *Earth Surf. Processes Landforms* **20**, 39–48 (1995).
- Temmerman, S., Goers, G., Wartel, S. & Meire, P. Spatial and temporal factors controlling short-term sedimentation in a salt and freshwater tidal marsh, Scheldt Estuary, Belgium, SW Netherlands. *Earth Surf. Processes Landforms* **28**, 739–755 (2003).
- Marion, C., Anthony, E. J. & Trentesaux, A. Short-term (≤ 2 yrs) estuarine mudflat and saltmarsh sedimentation: High-resolution data from ultrasonic altimetry, rod surface-elevation table, and filter traps. *Estuar. Coast. Shelf Sci.* **83**, 475–484 (2009).
- McKee, K. L., Cahoon, D. R. & Feller, I. C. Caribbean mangroves adjust to rising sea level through biotic controls on change in soil elevation. *Glob. Ecol. Biogeogr.* **16**, 545–556 (2007).
- Morris, J. T., Sundareshwar, P. V., Nietch, C. T., Kjerfve, B. & Cahoon, D. R. Responses of coastal wetlands to rising sea level. *Ecology* **83**, 2869–2877 (2002). **This article proposes that an optimum elevation (flooding frequency) for plant growth defines the transition from stable to unstable marsh.**
- Kirwan, M. L. & Guntenspergen, G. R. Feedbacks between inundation, root production, and shoot growth in a rapidly submerging brackish marsh. *J. Ecol.* **100**, 764–770 (2012).
- Marani, M., Lanzoni, S., Silvestri, S. & Rinaldo, A. Tidal landforms, patterns of halophytic vegetation and the fate of the lagoon of Venice. *J. Mar. Syst.* **51**, 191–210 (2004).
- Temmerman, S., Moonen, P., Schoelynck, J., Govers, G. & Bouma, T. J. Impact of vegetation die-off on spatial flow patterns over a tidal marsh. *Geophys. Res. Lett.* **39**, L03406 (2012).
- Möller, I. Quantifying saltmarsh vegetation and its effect on wave height dissipation: Results from a UK east coast saltmarsh. *Estuar. Coast. Shelf Sci.* **69**, 337–351 (2006).
- Mudd, S. M., D’Alpaos, A. & Morris, J. T. How does vegetation affect sedimentation on tidal marshes? Investigating particle capture and hydrodynamic controls on biologically mediated sedimentation. *J. Geophys. Res.* **115**, F03029 (2010).
- Cherry, J. A., McKee, K. L. & Grace, J. B. Elevated CO₂ enhances biological contributions to elevation change in coastal wetlands by offsetting stressors associated with sea-level rise. *J. Ecol.* **97**, 67–77 (2009).
- Kolker, A. S., Kirwan, M. L., Goodbred, S. L. & Cochran, J. K. Global climate changes recorded in coastal wetland sediments: empirical observation linked to theoretical predictions. *Geophys. Res. Lett.* **37**, L14706 (2010).
- Mariotti, G. & Fagherazzi, S. A numerical model for the coupled long-term evolution of salt marshes and tidal flats. *J. Geophys. Res.* **115**, F01004 (2010).
- Mariotti, G. *et al.* Influence of storm surges and sea level on shallow tidal basin erosive processes. *J. Geophys. Res.* **115**, C11012 (2010).
- Doyle, T. W., Krauss, K. W., Conner, W. H. & From, A. S. Predicting the retreat and migration of tidal forests along the northern Gulf of Mexico under sea-level rise. *For. Ecol. Manage.* **259**, 770–777 (2010).
- Morris, J. T., Edwards, J., Crooks, S. & Reyes, E. in *Recarbonization of the Biosphere: Ecosystems and the Global Carbon Cycle* (eds Lal, R. *et al.*) 517–531 (Springer, 2012).
- Yang, S. L., Milliman, J. D., Li, P. & Xu, K. 50,000 dams later: erosion of the Yangtze River and its delta. *Global Planet. Change* **75**, 14–20 (2011). **This article reports that upstream sediment restriction causes delta erosion that liberates enough sediment to sustain marshes.**
- D’Alpaos, A., Lanzoni, S., Marani, M. & Rinaldo, A. Landscape evolution in tidal embayments: modeling the interplay of erosion sedimentation and vegetation dynamics. *J. Geophys. Res.* **112**, F01008 (2007).
- Kirwan, M., Murray, A. & Boyd, W. Temporary vegetation disturbance as an explanation for permanent loss of tidal wetlands. *Geophys. Res. Lett.* **35**, L05403 (2008).
- Kearney, M. S., Rogers, A. S., Townsend, G., Rizzo, E. & Stutzer, D. Landsat imagery shows decline of coastal marshes in Chesapeake and Delaware Bays. *Eos* **83**, 173–178 (2002).
- Carniello, L., Defina, A. & D’Alpaos, L. Morphological evolution of the Venice lagoon: evidence from the past and trend for the future. *J. Geophys. Res.* **114**, F04002 (2009).
- Nyman, J. A., DeLaune, R. D., Roberts, H. H. & Patrick, W. H. Jr. Relationship between vegetation and soil formation in a rapidly submerging coastal marsh. *Mar. Ecol. Prog. Ser.* **96**, 269–279 (1993).
- Fagherazzi, S., Carniello, L., D’Alpaos, L. & Defina, A. Critical bifurcation of shallow microtidal landforms in tidal flats and salt marshes. *Proc. Natl Acad. Sci. USA* **103**, 8337–8341 (2006).
- Stevenson, J. C. & Kearney, M. S. in *Human Impacts on Salt Marshes: A Global Perspective* (eds Silliman, B. R., Grosholtz, E. D. & Bertness, M. D.) 171–206 (Univ. California Press, 2009).
- Davis, R. A., Yale, K. E., Pekala, J. M. & Hamilton, M. V. Barrier island stratigraphy and Holocene history of west-central Florida. *Mar. Geol.* **200**, 103–123 (2003).
- Balduff, D. M. *Pedogenesis, Inventory, and Utilization of Subaqueous Soils in Chincoteague Bay, Maryland*. PhD thesis, Univ. Maryland (2007).
- D’Alpaos, A., Da Lio, C. & Marani, M. Biogeomorphology of tidal landforms: physical and biological processes shaping the tidal landscape. *Ecohydrology* **5**, 550–562 (2012).
- Kirwan, M. L., Murray, A. B., Donnelly, J. P. & Corbett, D. R. Rapid wetland expansion during European settlement and its implication for marsh survival under modern sediment delivery rates. *Geology* **39**, 507–510 (2011).

44. Li, Y.-X., Törnqvist, T. E., Nevitt, J. M. & Kohl, B. Synchronizing a sea-level jump, final Lake Agassiz drainage, and abrupt cooling 8,200 years ago. *Earth Planet. Sci. Lett.* **315–316**, 41–50 (2012).
45. Cronin, T. M. *et al.* Rapid sea level rise and ice sheet response to 8,200-year climate event. *Geophys. Res. Lett.* **34**, L20603 (2007).
46. Donnelly, J. P. & Bertness, M. D. Rapid shoreward encroachment of salt marsh cordgrass in response to accelerated sea-level rise. *Proc. Natl Acad. Sci. USA* **98**, 14218–14223 (2001).
47. Kirwan, M. L. & Temmerman, S. Coastal marsh response to historical and future sea-level acceleration. *Quat. Sci. Rev.* **28**, 1801–1808 (2009).
48. Silliman, B. R. *et al.* Degradation and resilience in Louisiana salt marshes after the BP–Deepwater Horizon oil spill. *Proc. Natl Acad. Sci. USA* **109**, 11234–11239 (2012).
This article reports that vegetation mortality associated with oiling triggered rapid marsh edge erosion, and emphasizes the importance of vegetation health on marsh stability.
49. Smith, S. M. Multi-decadal changes in salt marshes of Cape Cod, MA: Photographic analyses of vegetation loss, species shifts, and geomorphic change. *Northeast. Nat.* **16**, 183–208 (2009).
50. Cahoon, D. R. *et al.* Mass tree mortality leads to mangrove peat collapse at Bay Islands, Honduras after Hurricane Mitch. *J. Ecol.* **91**, 1093–1105 (2003).
51. Silliman, B. R., van de Koppel, J., Bertness, M. D. & Mendelsohn, I. A. Drought, snails, and large-scale die-off of southern U.S. salt marshes. *Science* **310**, 1803–1806 (2005).
52. Alber, M., Swenson, E. M., Adamowicz, S. C. & Mendelsohn, I. A. Salt marsh dieback: an overview of recent events in the US. *Estuar. Coast. Shelf Sci.* **80**, 1–11 (2008).
53. Baustian, J. J., Mendelsohn, I. A. & Hester, M. W. Vegetation's importance in regulating surface elevation in a coastal salt marsh facing elevated rates of sea level rise. *Glob. Change Biol.* **18**, 3377–3382 (2012).
54. Langley, J. A. & Megonigal, J. P. Ecosystem response to elevated CO₂ levels limited by nitrogen-fuelled species shift. *Nature* **466**, 96–99 (2010).
This article reports that elevated CO₂ in isolation accelerated marsh elevation gain, but nitrogen additions caused a shift to a species unresponsive to elevated CO₂.
55. Langley, J. A., McKee, K. L., Cahoon, D. R., Cherry, J. A. & Megonigal, J. P. Elevated CO₂ stimulates marsh elevation gain, counterbalancing sea-level rise. *Proc. Natl Acad. Sci. USA* **106**, 6182–6186 (2009).
56. Bouillon, S. *et al.* Mangrove production and carbon sinks: a revision of global budget estimates. *Global Biogeochem. Cycles* **22**, GB2013 (2008).
57. Kirwan, M. L. & Blum, L. K. Enhanced decomposition offsets enhanced productivity and soil carbon accumulation in coastal wetlands responding to climate change. *Biogeochemistry* **8**, 987–993 (2011).
58. Kirwan, M. L. & Mudd, S. M. Response of salt-marsh carbon accumulation to climate change. *Nature* **489**, 550–553 (2012).
59. Charles, H. & Dukes, J. S. Effects of warming and altered precipitation on plant and nutrient dynamics of a New England salt marsh. *Ecol. Appl.* **19**, 1758–1773 (2009).
60. Gedan, K. B., Altieri, A. H. & Bertness, M. D. Uncertain future of New England salt marshes. *Mar. Ecol. Prog. Ser.* **434**, 229–237 (2011).
61. Beaumont, L. J. *et al.* Impacts of climate change on the world's most exceptional ecoregions. *Proc. Natl Acad. Sci. USA* **108**, 2306–2311 (2011).
62. McKee, K. L., Cahoon, D. R. & Feller, I. C. Caribbean mangroves adjust to rising sea level through biotic controls on change in soil elevation. *Glob. Ecol. Biogeogr.* **16**, 545–556 (2007).
63. Anisfeld, S. & Hill, T. D. Fertilization effects on elevation change and belowground carbon balance in a long island sound Tidal marsh. *Estuaries Coasts* **35**, 201–211 (2012).
64. Deegan, L. A. *et al.* Coastal eutrophication as a driver of marsh loss. *Nature* **490**, 388–392 (2012).
This article reports that long-term fertilization experiments led to channel expansion through decreased soil strength.
65. Turner, R. E. Beneath the salt marsh canopy: loss of soil strength with increasing nutrient loads. *Estuaries Coasts* **34**, 1084–1093 (2011).
66. Howes, N. C. *et al.* Hurricane-induced failure of low salinity wetlands. *Proc. Natl Acad. Sci. USA* **107**, 14014–14019 (2010).
67. Rooth, J. E. & Stevenson, J. C. Sediment deposition patterns in *Phragmites australis* communities: Implications for coastal areas threatened by rising sea-level. *Wetlands Ecol. Mgmt* **8**, 173–183 (2000).
68. Mozdzer, T. J. & Megonigal, J. P. Jack-and-master trait responses to elevated CO₂ and N: a comparison of native and introduced *Phragmites australis*. *PLoS ONE* **7**, e42794 (2012).
69. Nicholls, R. J. Coastal megacities and climate change. *GeoJournal* **37**, 369–379 (1995).
70. Syvitski, J. P. *et al.* Sinking deltas due to human activities. *Nature Geosci.* **2**, 681–686 (2009).
71. Törnqvist, T. E. *et al.* Mississippi Delta subsidence primarily caused by compaction of Holocene strata. *Nature Geosci.* **1**, 173–176 (2008).
72. Kolker, A. S., Allison, M. A. & Hameed, S. An evaluation of subsidence rates and sea-level variability in the northern Gulf of Mexico. *Geophys. Res. Lett.* **38**, L21404 (2011).
This article relates temporal trends in wetland loss to trends in subsidence rates and hydrocarbon extraction.
73. Turner, R. E. Wetland loss in the northern Gulf of Mexico: multiple working hypotheses. *Estuaries* **20**, 1–13 (1997).
74. Syvitski, J. P., Vorosmarty, C. J., Kettner, A. J. & Green, P. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. *Science* **308**, 376–380 (2005).
75. Tweel, A. W. & Turner, R. E. Watershed land use and river engineering drive wetland formation and loss in the Mississippi River birdfoot delta. *Limnol. Oceanogr.* **57**, 18–28 (2012).
76. Yang, S. L. *et al.* Impact of dams on Yangtze River sediment supply to the sea and delta intertidal wetland response. *J. Geophys. Res.* **110**, F03006 (2005).
77. Lotze, H. K. *et al.* Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science* **312**, 1806–1809 (2006).
78. Gedan, K. B., Silliman, B. R. & Bertness, M. D. Centuries of human-driven change in salt marsh ecosystems. *Annu. Rev. Mar. Sci.* **1**, 117–141 (2009).
79. Stedman, S. & Dahl, T. E. *Status and Trends of Wetlands in the Coastal Watersheds of the Eastern United States 1998–2004* (NOAA & US Department of the Interior, 2008).
80. Coleman, J. M., Huh, O. K. & Braud, D. Wetland loss in world deltas. *J. Coast. Res.* **24**, 1–14 (2008).
81. Giri, C. *et al.* Mangrove forest distributions and dynamics (1975–2005) of the tsunami-affected region of Asia. *J. Biogeogr.* **35**, 519–528 (2008).
82. Feagin, R. A. *et al.* Shelter from the storm? Use and misuse of coastal vegetation bioshields for managing natural disasters. *Conserv. Lett.* **3**, 1–11 (2010).
83. Das, S. & Vincent, J. R. Mangroves protected villages and reduced death toll during Indian super cyclone. *Proc. Natl Acad. Sci. USA* **106**, 7357–7360 (2009).
84. Barbier, E. B., Georgiou, I. Y., Enchelmeier, B. & Reed, D. J. The value of wetlands in protecting southeast Louisiana from hurricane storm surges. *PLoS ONE* **8**, e58715 (2013).
85. Barbier, E. B. *et al.* Coastal ecosystem-based management with nonlinear ecological functions and values. *Science* **319**, 321–323 (2008).
This article proposes that maximal economic value of mangrove forests can accommodate competing land uses.
86. Nicholls, R. J. Coastal flooding and wetland loss in the 21st century: changes under the SRES climate and socio-economic scenarios. *Glob. Environ. Change* **14**, 69–86 (2004).
87. van der Wal, D. & Pye, K. Patterns, rates and possible causes of saltmarsh erosion in the Greater Thames area (UK). *Geomorphology* **61**, 373–391 (2004).
88. Mattheus, C. R., Rodriguez, A. B., McKee, B. A. & Currin, C. A. Impact of land-use change and hard structures on the evolution of fringing marsh shorelines. *Estuar. Coast. Shelf Sci.* **88**, 365–376 (2010).
89. Siikamäki, J., Sanchirico, J. N. & Jardine, S. L. Global economic potential for reducing carbon dioxide emissions from mangrove loss. *Proc. Natl Acad. Sci. USA* **109**, 14369–14374 (2012).
90. Bauer, D. M., Cyr, N. A. & Swallow, S. K. Public preferences for compensatory mitigation of salt marsh losses: a contingent choice of alternatives. *Conserv. Biol.* **18**, 401–411 (2004).
91. Poulter, B., Qian, S. S. & Christensen, N. L. Jr. Determinants of coastal treelines, the role of abiotic and biotic interactions. *Plant Ecol.* **202**, 55–66 (2009).
92. Larsen, L. G. & Harvey, J. W. Modeling of hydroecological feedbacks predicts distinct classes of landscape pattern, process, and restoration potential in shallow aquatic ecosystems. *Geomorphology* **126**, 279–296 (2011).
93. Blum, M. D. & Roberts, H. H. Drowning of the Mississippi Delta due to insufficient sediment supply and global sea-level rise. *Nature Geosci.* **2**, 488–491 (2009).
94. Neubauer, S. C. Contributions of mineral and organic components of tidal freshwater marsh accretion. *Estuar. Coast. Shelf Sci.* **78**, 78–88 (2008).
95. Turner, R. E., Swenson, E. M. & Milan, C. S. in *Concepts and Controversies in Tidal Marsh Ecology* (eds Weinstein, M. & Kreeger, D. A.) 583–595 (Kluwer, 2000).
96. Schmidt, M. W. I. *et al.* Persistence of soil organic matter as an ecosystem property. *Science* **478**, 49–56 (2011).
97. Freeman, C., Ostle, N. & Kang, H. An enzymatic 'latch' on a global carbon store. *Nature* **409**, 149 (2001).
98. Megonigal, J. P., Hines, M. E. & Visscher, P. T. in *Anaerobic Metabolism: Linkages to Trace Gases and Aerobic Processes* (ed. Schlesinger, W. H.) 317–424 (Elsevier–Pergamon, 2004).
99. Craft, C. Freshwater input structures soil properties, vertical accretion, and nutrient accumulation of Georgia and U.S. tidal marshes. *Limnol. Oceanogr.* **52**, 1220–1230 (2007).
100. Weston, N. B., Vile, M. A., Neubauer, D. C. & Velinsky, D. J. Accelerated microbial organic matter mineralization following salt-water intrusion into tidal freshwater marsh soils. *Biogeochemistry* **102**, 135–151 (2011).

Acknowledgements The U.S.G.S. Global Change Research Program and the Virginia Coast Reserve Long Term Ecological Research Program (NSF DEB-0621014) supported this work financially. We thank G. Guntenspergen for conversations that enhanced this work.

Author Information Reprints and permissions information is available at www.nature.com/reprint. The authors declare no competing financial interest. Readers are welcome to comment on the online version of this article at go.nature.com/l7ijtf. Correspondence should be addressed to M.K. (kirwan@vims.edu).

The changing carbon cycle of the coastal ocean

James E. Bauer¹, Wei-Jun Cai², Peter A. Raymond³, Thomas S. Bianchi⁴, Charles S. Hopkinson⁵ & Pierre A. G. Regnier⁶

The carbon cycle of the coastal ocean is a dynamic component of the global carbon budget. But the diverse sources and sinks of carbon and their complex interactions in these waters remain poorly understood. Here we discuss the sources, exchanges and fates of carbon in the coastal ocean and how anthropogenic activities have altered the carbon cycle. Recent evidence suggests that the coastal ocean may have become a net sink for atmospheric carbon dioxide during post-industrial times. Continued human pressures in coastal zones will probably have an important impact on the future evolution of the coastal ocean's carbon budget.

The coastal ocean consists of several distinct but tightly connected ecosystems that include rivers, estuaries, tidal wetlands and the continental shelf. Carbon cycling in the coastal waters that connect terrestrial with oceanic systems is acknowledged to be a major component of global carbon cycles and budgets^{1–3}. Carbon fluxes within and between coastal subsystems, and their alteration by climate and anthropogenic changes, are substantial. It is therefore essential to understand, and accurately account for, the factors regulating these fluxes and how they affect the ocean and global carbon budgets.

Although the coastal contribution to the anthropogenic carbon-dioxide budget was neglected in past assessments reported by the Intergovernmental Panel on Climate Change (IPCC) and others, it has been recognized recently^{3–5}. Constraining the exchanges and fates of different forms of carbon in coastal settings has been challenging and is so far incomplete, due to the difficulty in scaling up relatively few observational studies. Rapid expansion of organic and inorganic carbon data collection (especially on the partial pressure of CO₂, *p*_{CO₂}) in coastal waters over the past decade, as well as new biogeochemical contexts for coastal systems dynamics, make this an exciting time for the field. A new generation of coupled hydrodynamic biogeochemical models can now mechanistically incorporate the factors that control carbon dynamics, such as elemental stoichiometry and biological turnover of both internally and externally supplied organic matter and nutrients, and their inputs and residence times^{6–8}. These new tools will provide a more predictive understanding of how coastal systems respond to human impacts and climate perturbations.

In this Review, we discuss our current understanding of the sources, fates and exchanges of organic and inorganic carbon in the coastal ocean, with an emphasis on the factors that contribute to net carbon fluxes within and between coastal subsystems. Carbon inputs and transformations are considered in the contexts of net air–water exchanges of CO₂, carbon burial in coastal subsystems, and exports to the open ocean. We explicitly address the growing recognition of how the coastal carbon cycle has fundamentally shifted in recent years owing to a variety of human activities. This synthesis shows that the present-day coastal ocean is a net sink for atmospheric CO₂ and a burial site for organic and inorganic carbon, and represents an important global zone of carbon transformation and sequestration. The purported shift of the coastal ocean from a CO₂ source to a CO₂ sink over the past 50 to 100 years also has ramifications for its future role in the ocean and global carbon cycles.

Riverine carbon inputs to coastal systems

Riverine supply of many elements, including carbon of largely terrestrial origin, is important to the steady-state chemistry of the oceans (Fig. 1a). Although estimates of riverine fluxes of both organic carbon^{9–11} and inorganic carbon^{3,12} continue to be improved by new geospatial tools and by scaling and modelling approaches, these fluxes are not greatly different from earlier estimates¹³ (Fig. 2 and Box 1). Average annual carbon fluxes to all major ocean basins and seas are now available^{10,12}. Fluxes are generally well correlated with river discharge, except in certain regions where factors such as high peat and carbonate coverage, and high erosion rates in watersheds also control carbon inputs^{10–14}.

Factors regulating riverine carbon fluxes

Climate has long been recognized as an important driver of river carbon supply to the coastal ocean (Fig. 1a). Watersheds with high precipitation have higher riverine discharge rates, and studies have long documented a primary regulation of carbon fluxes by discharge¹³, owing to the importance of transport limitation. Temperature also regulates important abiotic and biotic processes that can alter water throughput, flow paths, dissolution rates and watershed carbon stocks. The net effect of temperature on carbon fluxes can therefore vary between regions and among the different organic and inorganic forms of carbon (for example, see refs 15 and 16) (Fig. 1a).

In addition to annual precipitation and temperature, it is now clear that hydrologic 'events', such as extreme rainfall from tropical storms, are disproportionately important to riverine organic carbon transport. The erosive power of these storms is responsible for most particulate organic carbon (POC) export from watersheds to the coastal ocean, especially in mountainous regions¹⁷. Increases in riverine dissolved organic carbon (DOC) concentrations — and, hence, in annual riverine DOC export to coastal systems — can also result from these events. For example, a single tropical storm can be responsible for more than 40% of average annual riverine DOC export¹⁸. On decadal time scales, single large flood events can export 80–90% of POC from mountainous regions¹⁷. Climate models suggest that although the change in storm frequency is difficult to predict, the most intense storms will probably become more frequent⁴, and this will consequently affect riverine DOC and POC transport to coastal waters.

We can currently estimate with moderate to high certainty the riverine transport of terrestrial carbon to the coastal ocean (Fig. 2 and

¹Aquatic Biogeochemistry Laboratory, Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA. ²School of Marine Science and Policy, University of Delaware, Newark, Delaware 19716, USA. ³School of Forestry and Environmental Studies, Yale University, New Haven, Connecticut 06511, USA. ⁴Department of Geological Sciences, University of Florida, Gainesville, Florida 32611, USA. ⁵Department of Marine Science, University of Georgia, Athens, Georgia 30602, USA. ⁶Department of Earth & Environmental Sciences, Université Libre de Bruxelles, Brussels 1050, Belgium.

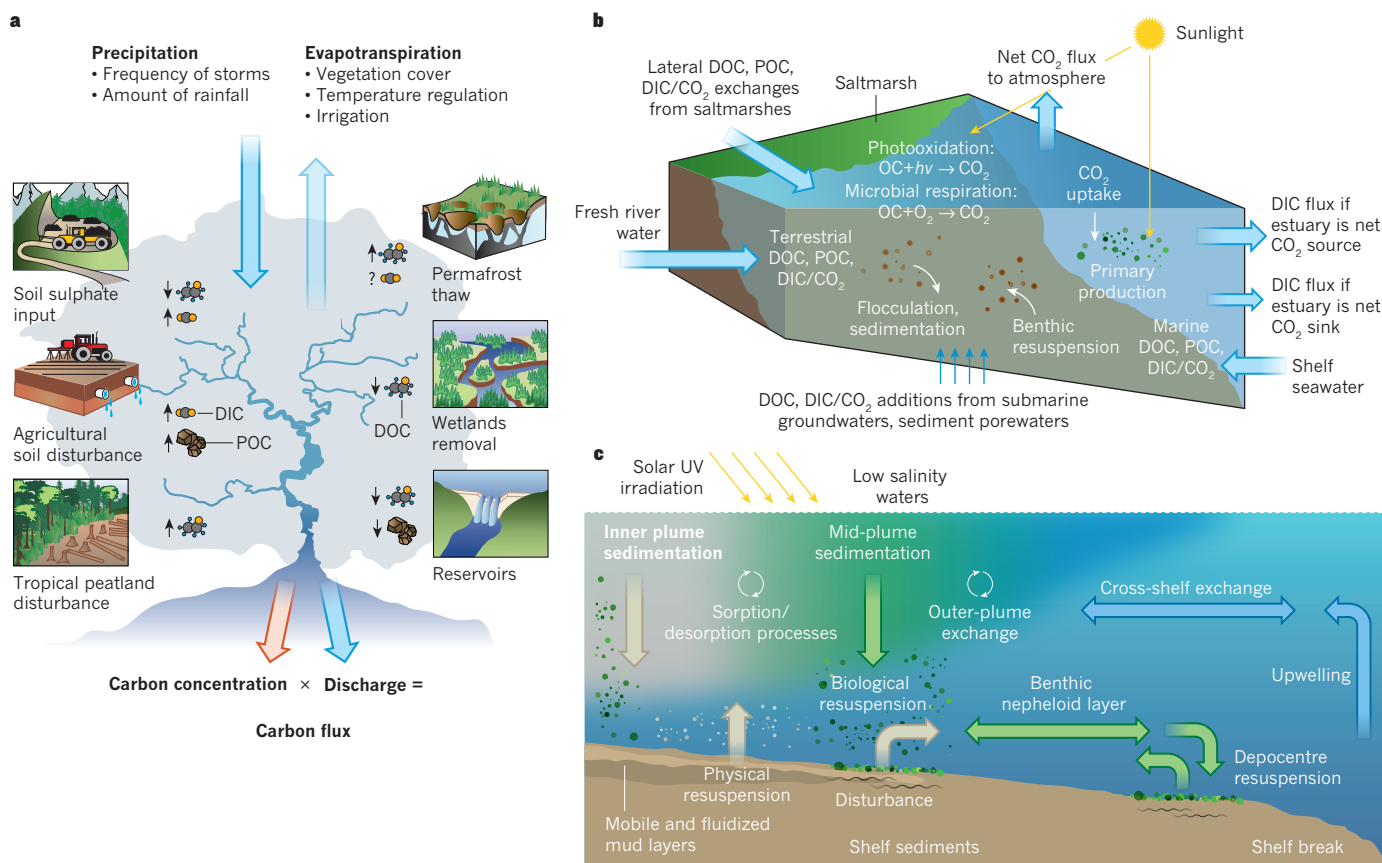


Figure 1 | Processes that affect organic and inorganic carbon cycling and fluxes in the major coastal ocean subsystems. **a**, Natural and anthropogenic processes altering riverine carbon inputs to the coastal ocean. Inputs can be altered through changes in the water balance (precipitation and evapotranspiration) and carbon stocks and flows in watersheds. Hydrological alterations include the effects of climate change on the amount and frequency of rainfall events and temperature regulation of evapotranspiration. Land management practices such as irrigation and clearance of vegetation that alter rates of evapotranspiration can also be important. Recent studies have indicated that inputs of sulphuric acid, agricultural practices, peatland disturbance, permafrost thaw, wetland removal and reservoir construction can alter carbon stocks (biogeochemical response) and flows through varied mechanisms at the drainage-network level. Carbon flux is equal to carbon concentration (the hydrological and biogeochemical response) multiplied by discharge (precipitation minus evapotranspiration). **b**, Major processes affecting carbon sources and fluxes in estuaries. Estuaries contain a mixture of organic and inorganic carbon sources derived from terrestrial materials carried by fresh river

Box 1). However, changing climate is expected to have major impacts on future river carbon fluxes and their uncertainties. Numerous studies now support precipitation as a dominant effect (compared with, for example, temperature) on these fluxes in the coming decades^{19,20}. Furthermore, rapid river transport times associated with large hydrological events will result in the bypassing of terrestrial carbon processing in rivers and concomitant episodic disturbance to coastal carbon budgets²¹, and lead to a shift in the timing of terrestrial carbon delivery to the coastal ocean under future climate change scenarios. The magnitude of these changes remains difficult to assess because the current generation of Earth system models does not simulate riverine carbon fluxes (Box 2).

Estuaries as modulators of carbon fluxes

Estuaries are transitional regions that range from predominantly river water to predominantly sea water. As a result, estuaries typically display strong gradients in biogeochemical parameters (for example, salinity, organic and inorganic carbon) during river and seawater mixing (Fig. 1b). Although

water (in which the salinity is zero), marine sources carried in shelf sea water (with a salinity of more than or equal to 30), and uniquely estuarine materials. Organic carbon is lost owing to salinity-induced flocculation, sedimentation, microbial respiration and photooxidation. Estuaries can modulate the export of carbon to the shelf depending on whether the estuary is a net carbon source or carbon sink. **c**, A representative continental shelf at its interface with a low-salinity river or estuarine plume. Physical and biogeochemical processes control the source, transport and fate of organic carbon. Carbon is exchanged at the interface between plume and shelf waters through sorption and desorption. Organic carbon transport to the open ocean is supplemented by physical resuspension, bioturbation and mobile and fluidized mud layers. The benthic nepheloid layer contains significant amounts of suspended sediment, which may be deposited to and resuspended from depocentres. Primary production in inner shelf waters may be limited by high sediment loads in plumes, whereas regions of upwelling in outer shelf waters can lead to elevated primary production. DOC, dissolved organic carbon; POC, particulate organic carbon; DIC, dissolved inorganic carbon.

most estuaries are geographically confined, estuaries of high-discharge rivers (for example, the Amazon) can extend onto, and even across, the continental shelf²². Numerous factors, such as the geomorphology of the estuary and the magnitude and stoichiometry of nutrient inputs (Box 3), control the fluxes and cycling of carbon in estuaries (Fig. 2). These important physical-biogeochemical reactors greatly modify the amounts and characteristics of organic and inorganic carbon transported between land and the ocean^{2,23,24}.

Organic carbon in estuaries

DOC and POC in estuaries are derived from terrestrial, marine and estuarine primary production (Fig. 1b). Owing to their unique biochemical and isotopic characteristics, specific sources of organic carbon have generally been easier to quantify than their fates within estuaries^{23,24}. *In situ* production of organic carbon in some estuarine waters can be significant to the coastal carbon budget because it can equal or exceed the river or marine supply^{25,26}.

Mineral sorption and desorption and photochemical dissolution can lead to an interchange between DOC and POC in estuaries^{27,28}.

Estuaries also experience significant losses of organic carbon owing to the combined influences of microbial degradation and photochemical oxidation^{29–31}, scavenging, sedimentation, and salinity-induced flocculation of DOC and POC³² (Fig. 1b). Individual sources of organic carbon have unique reactivity and residence times that affect their degradation⁶ to climatically important gases such as CO₂, methane and volatile organic carbon in estuarine waters and sediments³³, and their export to continental shelves. Estuaries can therefore modulate organic carbon exports to shelves relative to riverine organic carbon fluxes alone³⁴.

Most estuaries have tremendous internal spatial and temporal heterogeneity in carbon processing and fluxes, making it difficult to quantify even a single estuary's net carbon balance. As a result, we lack the measurements from a representative number of systems for accurate global or even regional estimates of the direction and magnitude of net organic carbon fluxes that occur within estuaries. In addition, the complex interplay between organic carbon and both inorganic and organic nutrient inputs (Box 3) from land and ocean have an important, but poorly quantified, role in regulating the balance between net organic carbon production (autotrophy) and consumption (heterotrophy) in estuaries^{35,36}. Process-based models of coupled estuarine hydrodynamics and biogeochemistry have recently addressed interactions between the organic and inorganic carbon cycles at the scale of individual estuaries (for example, see refs 8, 21, 35, 37), but none are currently suitable for regional or global applications (Box 2).

Estuarine carbon dioxide and inorganic carbon exchange

CO₂ emissions from European estuaries were recognized to be a significant component of the regional CO₂ budget³⁸ about 15 years ago. Subsequent studies estimated global estuarine CO₂ emissions to be on the order

of 0.2–0.4 Pg C yr⁻¹ (refs 2, 33, 39, 40). Estuaries occupy a small portion of global ocean area (about 0.2%), and, therefore, their CO₂ emissions are a disproportionately large flux when compared with CO₂ exchanges between the open ocean and atmosphere⁴¹ (Fig. 3a). However, the uncertainty in the global estuarine CO₂ emission flux is high (Fig. 2, 3a,b and Box 1) due to very limited spatial and temporal coverage during field observations, large physical and biogeochemical variability and insufficient use of generalized hydrodynamic–biogeochemical models in estuaries^{2,22}.

In addition to very high p_{CO_2} and correspondingly high rates of CO₂ degassing, dissolved inorganic carbon (DIC) is also usually enriched in estuarine waters and exported to continental shelf waters (Fig. 2). The elevated p_{CO_2} and DIC result from *in situ* net respiration of internally and externally supplied organic carbon and lateral transport of DIC from rivers and coastal wetlands^{2,33,40,42,43} and CO₂-rich groundwaters⁴⁴. Additionally, low-DIC estuaries⁴² generally experience higher CO₂ degassing fluxes than high-DIC estuaries, as waters of the latter retain CO₂ longer owing to their greater buffering capacity^{45,46}.

Exchanges with tidal wetlands

Highly productive tidal wetlands flank many estuaries and laterally export both dissolved and particulate carbon to estuaries and coastal systems^{2,47}. Regional and global estimates of wetland fluxes are hampered by a scarcity of reliable estimates of wetland surface area and studies of carbon export. The limited studies that are available suggest that wetlands act as a net source of carbon to estuaries that could be comparable with riverine carbon supply^{2,3} (Fig. 2). However, although some of the carbon exported from wetlands is recycled in estuaries, significant amounts are buried in estuarine sediments and exported to continental shelves (Fig. 2).

Globally, estuaries are net heterotrophic, meaning that respired organic carbon exceeds that supplied by rivers and wetlands, and produced *in*

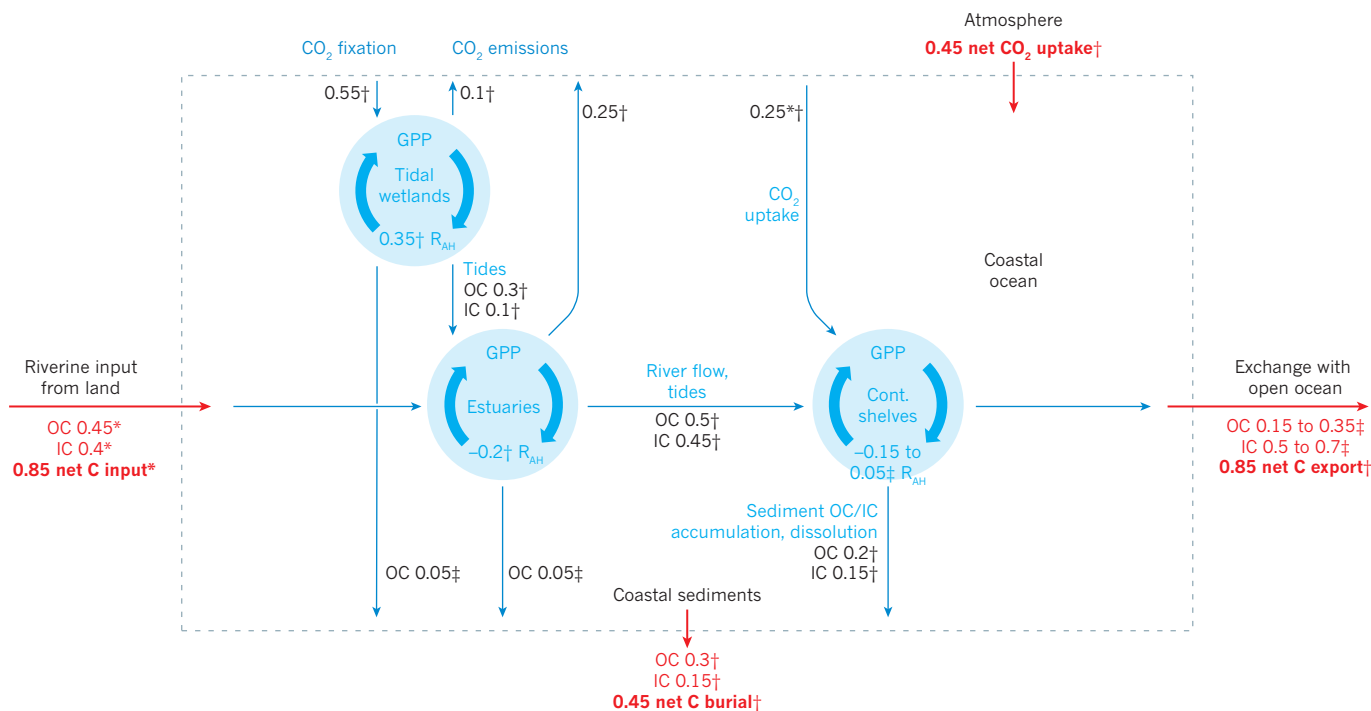


Figure 2 | Organic and inorganic carbon fluxes in the estuarine, tidal wetland and continental shelf subsystems of the coastal ocean. Fluxes between adjacent subsystems and other components of the earth system are regulated by a number of processes (the major ones are shown here). Carbon can flux both within (values in black) and across (values in red) the boundaries of the coastal ocean. All organic carbon (OC) and inorganic carbon (IC) fluxes are presented as positive values, arrows indicate direction of flux. Particulate and dissolved OC fluxes are presented as total OC values. The balance between gross primary production (GPP) and total system

respiration (both autotrophic, A, and heterotrophic, H; R_{AH}) is net ecosystem production (NEP), with negative values indicating conversion of OC to IC. The IC burial flux takes into consideration calcification. The methods used to estimate flux values and their associated uncertainties are described in Box 1. Typical uncertainties for carbon fluxes: *95% certainty that the estimate is within 50% of the reported value; †95% certainty that the estimate is within 100% of the reported value; ‡uncertainty greater than 100%. Units are Pg C yr⁻¹ (1 Pg = 10¹⁵ g) rounded to ±0.05 Pg C yr⁻¹. Within-river fluxes and transformation of carbon are excluded from this analysis.

BOX 1

Coastal carbon flux estimates and their uncertainties

Carbon fluxes and their associated uncertainty estimates presented in this Review are not based on statistical treatment of multiple observed data, as data coverage is poor and often skewed. Rather, most fluxes are based on ranges presented in the literature and the quality of the individual values. Because of the large degree of heterogeneity in the main coastal subsystems and concomitant lack of data, most carbon fluxes in these subsystems have relatively high uncertainties. Riverine carbon fluxes have been estimated over the past three decades and are known with the highest degree of confidence (95% certainty that the estimate is within 50% of the reported value^{2,3,9–14}) (Fig. 2). Wetland net primary production rate, carbon burial, CO₂ degassing and dissolved inorganic carbon (DIC) export rates are based on values in the literature^{2,49,86}. The estuarine CO₂ degassing flux is based on the best known syntheses of field data^{2,33,40,66} with moderate confidence (95% certainty that the estimate is within 100% of the reported value) (Figs 2, 3). Organic carbon burial of 0.03–0.2 Pg C yr⁻¹ in wetland and estuarine sediments is the most poorly constrained flux (95% certainty that the estimate is greater than 100% of the reported value), and we adopt a conservative value^{2,3} (Fig. 2). Lateral carbon fluxes from estuaries to continental shelves are estimated from the estuarine mass balance of river and wetland inputs, estuarine CO₂ degassing, and marsh and estuarine net ecosystem productivities (NEP), and therefore have moderate confidence.

Owing to recent marked improvements in the quantity and quality of data, present-day continental shelf CO₂ gas exchange (0.25 Pg C yr⁻¹)

is among the best estimated of all coastal carbon fluxes (within 50% uncertainty)^{2,39,40,61,66}. However, because of less-constrained fluxes in enclosed seas and low latitude open shelves (up to 75% uncertainty)²² (Fig. 3c), we assign an overall uncertainty for shelves of 50–75%. On the basis of the CO₂ uptake flux, the known global shelf surface area (26×10^6 km²), average gas transport parameter (9.3 cm h⁻¹), and present-day atmospheric p_{CO_2} (Fig. 4b), an average surface water p_{CO_2} of 350 ± 18 p.p.m. is estimated for the post-industrial shelf. Present-day carbon burial in shelf sediments is also estimated with moderate confidence^{2,3,50,89}.

For the pre-industrial shelf (Fig. 4a) a net CO₂ degassing flux of around 0.15 Pg C yr⁻¹ was estimated as the mean of the upper^{2,50} and lower³ bounds of reported values. To be compatible with this CO₂ degassing flux and a pre-industrial atmospheric p_{CO_2} of 280 p.p.m., an average surface water p_{CO_2} of 298 ± 18 p.p.m. is estimated (Fig. 4a). For pre-industrial time, we estimate a shelf NEP of -0.15 Pg C yr⁻¹ as the mean between an upper limit, assuming that 60% of terrestrial organic carbon is respired on shelves and a low bound of zero from ref. 68. However, for present-day shelf NEP, we assumed two scenarios (see 'Pre- and post-industrial shelf carbon budget'). For the pre-industrial and present-day comparison of shelf CO₂ fluxes, we held constant river input, estuarine CO₂ exchange flux with the atmosphere, NEP and sediment burial constant and then calculated the various organic carbon and inorganic carbon export fluxes by mass balance. Our level of confidence for these estimates is low (Fig. 4a, b).

situ, by 0.2 Pg C yr⁻¹ (ref. 33). The outgassing of estuarine CO₂, derived from organic carbon respiration and DIC inputs from rivers and wetlands, releases around 0.25 Pg C yr⁻¹ of CO₂ into the atmosphere (Figs 2, 3a, b). Although highly uncertain, carbon burial in wetlands and estuaries is probably around 0.1 Pg C yr⁻¹ (refs 3, 48, 49) (Fig. 2 and Box 1). Our mass balance analysis suggests that estuaries export about 10% more carbon to continental shelves than they import from rivers (Fig. 2).

Tremendous geomorphological differences between estuaries and a lack of synthetic modelling lead to considerable variability and uncertainty in estimates of estuarine carbon dynamics and export to continental shelves. Estuaries are typically net sources of CO₂ to the atmosphere and augment the organic and inorganic carbon supply to shelves (Fig. 2). Climate and land-use changes (specifically sea-level rise and declining river sediment export) are likely to decrease net carbon burial in estuaries and wetlands⁴⁹; however, with anticipated increased inputs from rivers, estuarine export of carbon to continental shelves is likely to increase.

Carbon on continental shelves

Continental shelves are dynamic interfaces where terrestrial, estuarine and marine organic carbon is recycled (Fig. 1c). Continental shelves occupy only 7–10% of global ocean area^{22,50} (Fig. 3a). However, shelves contribute 10–30% of global marine primary production, 30–50% of inorganic carbon and around 80% of organic carbon burial in sediments⁵⁰, and could contribute up to about 50% of the organic carbon supplied to the deep open ocean⁵¹. Thus, continental shelves are disproportionately important to ocean carbon cycles and budgets.

Shelf organic carbon sources and sinks

Continental shelf primary production is often related to shelf width and the magnitude of river discharge⁵² (Fig. 1c). On broad river-dominated shelves (for example, Mississippi and Yangtze), primary production on the inner shelf may be limited by high particulate loads, and inputs of river- and estuary-derived organic carbon may dominate the water column and sediments. On broad shelves with lower river discharge (for example,

South Atlantic Bight), sunlight may reach the sea floor and support a significant benthic contribution to shelf primary production and organic carbon⁵¹. On narrow shelves, pelagic and/or benthic primary production is supported by oceanic inputs of nutrients and is recycled on the shelf⁶².

Organic carbon burial rates on continental shelves are controlled by multiple mechanisms (Fig. 1c). When river- and estuarine-derived organic carbon enters the coastal zone, about 90% of it is associated with mineral matrices in organo-clay aggregates²⁴. In many highly productive upwelling regions along shelves, organic carbon particles may aggregate by glue-like exopolymers from phytoplankton⁵³. In addition to burial, organic carbon in shelf environments is transported to the open ocean through physical resuspension, bioturbation, and mobile and fluid muds along the sea floor⁵⁴ (Fig. 1c). Flocculation processes similar to those in estuaries (Fig. 1b) can also be important in transporting selected forms of DOC and POC from shelf waters to sediments and altering their chemical composition⁵¹.

Organic carbon reactivity on shelves

Although terrestrial DOC supplied by rivers can theoretically account for the 4,000–6,000 year residence time of DOC in the global ocean⁵⁵, little terrestrial material is actually detected in the oceans. How this large flux of terrestrial DOC is processed in the coastal ocean is a major remaining question in coastal carbon research. Recent findings show that much (90% or more) of the reactive aromatic fraction of terrestrial DOC is altered in inner shelf waters by sunlight-driven photoreactions to produce highly stable, ubiquitous and presumably long-lived components of oceanic DOC^{56,57}. Bacterial alteration of terrestrial plant-derived lignin can account for additional DOC losses (up to 30% of the photochemical losses) in river-dominated shelf waters⁵⁸. There is, thus, a growing consensus that much of the terrestrial DOC becomes chemically altered, rather than completely oxidized to CO₂, on shelves. This may help to reconcile the disagreement between riverine inputs and ocean DOC residence times. In contrast to terrestrial organic carbon, 50–90% of the organic carbon that is derived from marine primary production is rapidly recycled

on continental shelves. Reactive marine material may also enhance the metabolism of less reactive terrestrial organic carbon in shelf waters and sediments⁵⁹.

Carbon–dioxide exchange in shelf waters

Tsunogai *et al.*⁶⁰ first pointed out, in 1999, the importance of continental shelf CO₂ uptake to the carbon cycling and climate change communities, proposing a value as high as around 1 Pg C yr⁻¹, equal to 50% of the open ocean CO₂ uptake known at the time. Most recent syntheses are based on up-scaling methods whereby different shelf systems are classified by dividing them into a few provinces⁶¹ or typologies⁴⁰. These estimates suggest a lower, but, relative to the global ocean and land⁴¹, still significant net atmospheric CO₂ uptake flux of 0.25 Pg C yr⁻¹ (Fig. 3a).

Inner continental shelf waters close to land tend to be sources of CO₂, mostly due to their high rates of respiration of terrestrial and estuarine organic carbon and lateral transport of high CO₂ waters from adjacent inshore systems. By contrast, mid- to outer-shelf waters are a sink of CO₂ (ref. 62). This general pattern results from decreased terrestrial organic carbon supply, increased primary production as light conditions improve offshore, and increased accessibility to nutrients supplied by upwelling and mixing across the shelf break⁶³. This pattern and the inshore to offshore shift from CO₂ release to CO₂ uptake across shelves can be altered greatly in larger river plumes⁶⁴ or in upwelling dominated shelves⁶⁵ (Fig. 1c) and depends to a significant extent on physical conditions such as wind stress and river discharge. Furthermore, a striking latitudinal contrast in shelf-water–atmosphere CO₂ fluxes emerges from a global synthesis of shelf systems. Present-day shelves located between 30° and 90° latitude are, in general, sinks for atmospheric CO₂ whereas shelves located between the equator and 30° tend to be sources of CO₂ (or nearly neutral) to the atmosphere^{61,66} (Fig. 3c). This latitudinal pattern could, in part, be explained by the fact that around 60% of river organic carbon is exported to lower latitude shelves and respired under the higher mean temperatures in these systems^{2,10}. Of increasing importance is that the western Arctic Ocean margin, in particular the nutrient-rich Chukchi Sea, has become a rapidly increasing global shelf

sink for atmospheric CO₂ over the past decade⁶⁷ due to greater annual retreat of sea ice — a major barrier to gas exchange — as a result of climate warming.

Autotrophic–heterotrophic balance

Before the extensive alteration of terrestrial landscapes and industrial fertilizer production, estuarine and continental shelf waters were on the whole thought to be net heterotrophic (they released more CO₂ to the atmosphere than was fixed by primary production, as a result of their respiration of terrestrial and tidal wetland organic carbon inputs^{68,69}). Indeed, almost every river and estuary globally, for which data are available, is today a strong source of CO₂ (refs 1, 40, 66). These systems were probably an equally strong — if not stronger — source of CO₂ in the past when atmospheric p_{CO_2} was lower.

There is currently no consensus as to whether present-day continental shelves are net autotrophic or heterotrophic, owing, in part, to a lack of concurrent respiration measurements to go along with the abundant primary productivity measurements that have been made in shelf waters⁷⁰. The complexity of coastal systems further hampers proper upscaling for modelling (Box 2). A school of thought suggests that continental shelves as a whole are now net autotrophic because of increased anthropogenic nutrient supply^{50,63,69}, which in some systems exceeds deep ocean nutrient inputs by upwelling or mixing across the shelf break⁶³. This view is consistent with the observation that most shelves are a net sink for atmospheric CO₂ (refs 40, 61, 66). Thus, it has been postulated that the coastal ocean as a whole has shifted from a net heterotrophic to an increasingly net autotrophic state, which has in turn favoured a reversal from shelves being a CO₂ source to a CO₂ sink in recent decades⁶⁹. This hypothesis is substantiated by results from coarse-grained box models (Box 2). Although it is important to establish whether such a shift has occurred, if it has, its exact magnitude and timing remain highly uncertain.

Similar to estuaries, continental shelves are highly heterogeneous coastal subsystems. Carbon dynamics in some shelves are controlled entirely by ocean circulation, whereas in others they are controlled largely by riverine inputs. Uncertainties in shelf carbon fluxes are significant (Fig. 2) but cannot yet be adequately constrained

BOX 2

Modelling the coastal carbon cycle

Earth system models (ESMs) are climate models that can include physical processes, as well as biogeochemical cycles, and which allow a representation of anthropogenic processes⁹⁵. They describe processes within and between the atmosphere, ocean, cryosphere, and terrestrial and marine biosphere. ESMs include coarse-grained box models, models of intermediate complexity and comprehensive tridimensional global climate models that incorporate biogeochemical processes, such as carbon cycle and atmospheric chemistry. However, with the exception of a few box models, ESMs are at present limited by their lack of coupling between atmosphere, land and ocean components through lateral flows of carbon (and nutrients; Box 3) along the land–ocean continuum³. The delivery of riverine carbon is included as a forcing condition in large-scale ocean component models⁹⁶, but spatially resolved ESMs do not simulate the riverine carbon fluxes dynamically.

Attempts to estimate the historical evolution of the aquatic fluxes from land to ocean and their effects on estuarine and continental shelf carbon dynamics have so far relied solely on globally averaged box models⁶⁹. Although these models are extremely valuable for testing further conceptual ideas, they rely on highly parameterized process formulations. In addition, global box models do not account for the wide diversity of estuarine and shelf systems, nor do they mechanistically represent the effect of land-use changes on terrestrial biogeochemistry

and riverine fluxes.

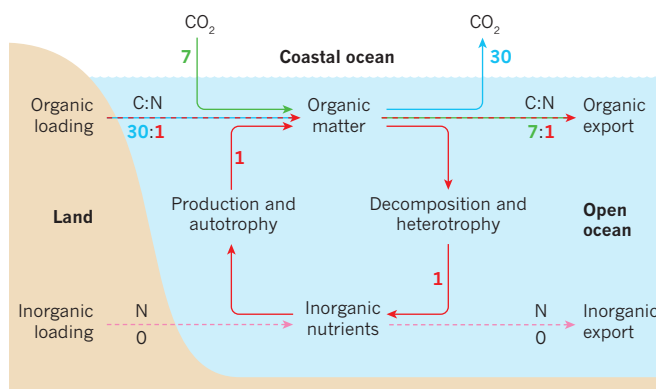
State-of-the-art ESMs contain routing schemes for riverine water flows at a spatial resolution of 0.5°. At this resolution, about 500 rivers that contribute around 80% of the total organic carbon delivery to the oceans may now be resolved. Furthermore, with the development of eddy-permitting runs at a resolution of about 0.25°, shelf carbon dynamics are increasingly better captured. Including these dynamics will allow better resolution of important processes such as the air–sea CO₂ exchange that results from large-scale coastal eddies or the physical transport divergence of carbon across coastal boundaries and the shelf break. The largest rivers of the world produce coastal plumes which can also be reasonably well captured⁹⁷, whereas narrow coastal systems such as eastern boundary currents will still require higher resolution⁹⁸. For estuaries and tidal wetlands a resolution of 0.25–0.5° is too coarse, and specific modelling approaches that rest on mechanistically rooted upscaling strategies need to be designed to better constrain their roles in ocean and global carbon cycles and assess their sensitivity to anthropogenic disturbances. Improved mechanistic descriptions of ecosystem and biogeochemical processes are also needed as the current formulations may not be able to sufficiently capture the complexity of coastal ocean dynamics and their response to enhanced terrestrial inputs³.

BOX 3

The role of elemental stoichiometry

Coastal subsystems tend to release CO_2 to the atmosphere simply because of the contrasting elemental stoichiometry (the ratio of carbon to nitrogen, C:N) of terrestrial organic matter (TOM) decomposition and aquatic primary production. TOM carried to the coast by rivers typically has a high C:N (30:1 to 60:1) (Box Fig., left), whereas organic matter produced by algae such as phytoplankton in coastal systems has C:N near 7:1 (Box Fig., right).

The stoichiometry of decomposition and production crucially influences the metabolic balance of aquatic systems (the ratio of gross primary production (GPP) to ecosystem respiration (R)). During TOM decomposition, 30–60 moles of inorganic carbon (that is, CO_2) are produced for every 1 mole of inorganic nitrogen produced. If 100% of the remineralized nitrogen is recycled and taken up by algal primary production, only 7 moles of inorganic carbon can be taken up by new algal biomass having a C:N stoichiometry of 7:1. That is, relative to nitrogen, much more CO_2 is produced during decomposition than is taken up during primary production. Thus, coastal systems with significant decomposition of high C:N TOM will tend to be heterotrophic, with respiration (production of CO_2) exceeding primary production (consumption of CO_2). Equally important in controlling the net uptake or release of CO_2 are the rates of TOM decomposition (days to years) compared with algal



organic matter decomposition (hours to months) and algal primary production (hours to days) relative to water residence time^{99,100}. This mismatch in time scales of autotrophic and heterotrophic processes can drive the net metabolic balance³⁵. The system shown in the figure is net heterotrophic ($P < R$) and NEP is -23 ($P - R$). Nitrogen loaded from land and recycled is shown in red, terrestrial carbon is blue and recycled production is green.

through statistical treatments of available data sets (Box 1). These uncertainties represent an important barrier for the integration of coastal systems in global carbon-cycle assessments. Although the magnitude of each flux may be revised in further assessments such that uncertainties are reduced, this is unlikely to affect our analysis that shelf–atmosphere CO_2 exchanges and lateral carbon fluxes on shelves are important to the global carbon balance. Changes in riverine and estuarine organic carbon and nutrient supply and increasing atmospheric levels of CO_2 are all postulated to alter shelf–atmosphere–open-ocean exchanges.

Human impacts on the coastal carbon cycle

Major uncertainties remain in our understanding of the natural, undisturbed carbon cycle of the coastal ocean. Activities such as land-use modification, waterway impoundment, nutrient inputs, wetland degradation and climate change add even greater complexity and uncertainty, making it difficult to differentiate the natural and anthropogenic drivers affecting changes in the coastal carbon cycle. Although a number of drivers of anthropogenic disturbance to river and estuarine carbon fluxes have been identified, quantitative estimates of their effect on these fluxes are very poorly constrained. Understanding the exact direction and magnitude of different human impacts on individual fluxes is not only important in both coastal and global carbon budgets but also for determining the global terrestrial CO_2 sink — a flux defined as the residual from all other components of the anthropogenic CO_2 budget^{3,71}. Research on riverine organic carbon export, in particular, shows that at least part of the anthropogenic CO_2 taken up by land ecosystems is exported to the coastal ocean^{3,4}. The change in land carbon storage as synthesized by the IPCC⁴ and others may thus be overestimated because a significant fraction of the displaced carbon is stored in coastal waters and sediments³. Quantification of the temporal evolution of coastal carbon transfers and fluxes, and the incorporation of coastal carbon processes in Earth system models is also necessary to support relevant policies and mitigation strategies (Box 2).

Human perturbations to rivers and estuaries

Land management is now considered to be a primary driver for changing riverine carbon exports to the coastal ocean (Fig. 1a). Agricultural

practices have led to an increase in the movement of sediment and POC from land⁷². This does not necessarily lead to increased carbon fluxes to the coastal ocean, however, because much of the terrestrial material is re-deposited on land⁷³ or trapped in man-made reservoirs and agricultural impoundments¹¹. For example, reservoirs are thought to have reduced current POC fluxes to around 90% of the pre-anthropogenic level¹¹. Similarly, agriculturally enhanced riverine fluxes of bicarbonate and other major ions are now reported in numerous regions and have increased by as much as 40% in systems such as the Mississippi, owing to liming and hydrological alterations⁷⁴. In addition, organic carbon exported by agricultural landscapes may have different chemical characteristics, and lower overall susceptibility to photochemical and microbial degradation than organic carbon from less altered or forested landscapes⁷⁵. Conversion of wetland and peatland systems to other land uses can also affect the amounts and ages of organic carbon transported to rivers and estuaries^{76,77}. Newer spatially resolved models that include riverine fluxes at the scale of large watersheds suggest increased fluxes, owing to CO_2 stimulation of terrestrial primary production under future climate scenarios¹⁵.

There is growing evidence that humans have altered the riverine flux of carbon to the coastal ocean on regional to global scales. Regional increases in riverine DOC concentration have been reported, with multiple mechanisms suggested⁷⁸. It is also increasingly likely that global DIC fluxes have increased as a result of liming and anthropogenic acid additions to watersheds^{74,79}, and that POC fluxes have decreased owing to changing precipitation regimes and land and river management. Although future climate change is predicted to lead to an increase in river carbon fluxes (Fig. 1a), it is also likely to lead to increased uncertainties in predicting these fluxes.

Human activities associated with estuaries, such as land-use change, bulkhead construction, wastewater discharge, wetland removal and dredging, are increasingly important factors that affect estuarine carbon sources, cycling and budgets³ (Fig. 1b). Elevated levels of organic carbon respiration in productive, nutrient-rich estuarine waters and river plumes have recently been found to amplify the effects of ocean acidification on estuarine ecosystems^{80,81}. In addition, the microbial oxidation of ammonia and organic carbon from urban wastewater discharge generates acids, leading to very low pH and high p_{CO_2} conditions and enhancing

CO₂ degassing in some estuarine waters^{31,82}. Changes in physical drivers such as river discharge, sea-level rise or storm frequency and intensity may fundamentally alter estuarine biogeochemical processes (Box 3) and consequently CO₂ and DIC exchanges between estuaries, the atmosphere and continental shelves^{3,83} (Fig. 1b). Continued environmental changes in estuaries and associated wetlands will have potential impacts on the direction and magnitude of estuarine internal organic and inorganic sources and sinks, and the magnitude of both their atmospheric CO₂ emissions and their export of organic and inorganic carbon to shelves.

Human impacts on shelf organic carbon

Human impacts on processes such as river discharge, nutrient inputs and climate warming have also resulted in changes to continental shelf organic and inorganic carbon cycling. Terrestrially derived DOC from soils and thawing high-latitude peatlands may limit light penetration and primary production in continental shelf waters (Fig. 1c), and its degradation by sunlight can result in production of radiatively important trace gases such as CO₂, CO and CH₄. These concerns are compounded in the environmentally sensitive Arctic Ocean, where recent marked decreases in summer time sea-ice cover are hypothesized to lead to an increasing role for photochemistry in Arctic shelf organic carbon cycling⁸⁴. In addition, recent reductions in POC loads by river damming globally (Fig. 1a) have resulted in relatively greater contributions of estuarine-derived organic carbon⁵⁹, which is generally considered to be more biologically reactive than terrestrial organic carbon. Increased terrestrial export of nutrients and carbon by rivers due to agricultural activities and land-use change has also led to excess primary production and organic carbon accumulation in estuarine and shelf waters and sediments, which in turn may lead to oxygen-depleted hypoxic conditions — now considered to be a major global environmental issue in temperate and tropical coastal systems⁸⁵.

Pre- and post-industrial shelf carbon budget

The continental shelf was probably net heterotrophic during pre-industrial times, because a large fraction of terrestrial organic carbon export must have been respired in shelf waters^{2,50,68,69}. In the following analysis, we assume a pre-industrial shelf net ecosystem production (NEP) of $-0.15 \text{ Pg C yr}^{-1}$ (a mean between 60% of terrestrial organic carbon respiration and shelf NEP of zero⁶⁸) and a net shelf CO₂ degassing flux of about $0.15 \text{ Pg C yr}^{-1}$ (a mean between a high bound^{50,68} and a low bound³) (Fig. 4a). This CO₂ degassing flux, together with the known global shelf area, average gas transport parameter and pre-industrial atmospheric p_{CO_2} (Fig. 4a), requires an average pre-industrial surface water p_{CO_2} of $298 \pm 18 \text{ ppm}$ (Box 1). Our mass balance analysis also estimates moderate DIC and DOC export fluxes to the ocean during pre-industrial times (Fig. 4a). The combined pre-industrial offshore DIC and organic carbon fluxes ($0.45 \text{ Pg C yr}^{-1}$) estimated here (Fig. 4a) are therefore much less than the riverine or estuarine DIC and organic carbon fluxes ($0.95 \text{ Pg C yr}^{-1}$) (Fig. 2) because of extensive CO₂ outgassing and burial.

A drastically different scenario emerges for the post-industrial period, when shelves are at present a CO₂ sink and calculated to have an average global shelf p_{CO_2} of $350 \pm 18 \text{ ppm}$ under present-day atmospheric p_{CO_2} of 380 p.p.m. (Fig. 4b and Box 1). Two mechanisms may be responsible for the shelf switching from a CO₂ source to a CO₂ sink between pre- and post-industrial times. The prevailing view assumes a shift in continental shelf NEP over the industrial age from net heterotrophic to increasingly net autotrophic conditions (NEP up to $+0.05 \text{ Pg C yr}^{-1}$). This is a result of enhanced biological uptake of atmospheric CO₂ owing to stimulation of shelf primary production by increased anthropogenic nutrient inputs from land⁶⁹. This mechanism would contribute to the reversal in shelf CO₂ flux over the industrial era (Box 3) and implies an increase in organic carbon export from shelves to the open ocean to about $0.35 \text{ Pg C yr}^{-1}$ (Fig. 4b) or higher⁸⁶ and a moderate increase in DIC export to about 0.5 Pg C yr^{-1} .

Here we propose an alternative mechanism that could also explain the present-day continental shelf CO₂ sink. This mechanism requires no change in shelf NEP but simply an increased physical uptake of atmospheric CO₂, as atmospheric levels of CO₂ have risen to a much greater

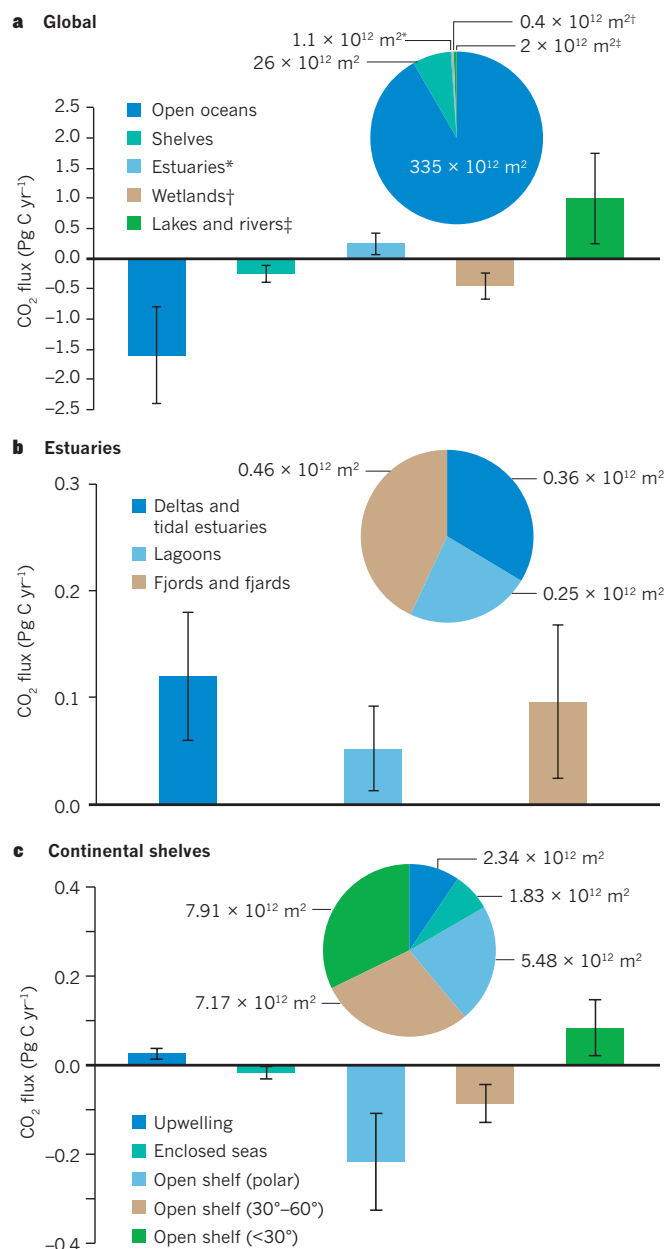


Figure 3 | Air-surface water CO₂ exchange fluxes of different aquatic systems. **a**, The global areas of major aquatic systems. **b**, The CO₂ flux and global areas of river deltas and embayments (estuary type I and II in refs 33 and 40), coastal lagoons (type III), and fjords and fjards (type IV) (note that large river plumes extending onto and across continental shelves are not included, although it is known that they are a CO₂ sink for the atmosphere). **c**, Carbon flux and global areas of continental shelves. Continental shelves freely exchange with the open ocean in low latitudes (0–30°), temperate or mid-latitudes (30–60°) and polar or high latitudes (60–90°), upwelling systems, and enclosed shelf seas. We assigned an uncertainty of <50% to each flux term except inland waters, fjords and fjards, enclosed seas and low latitude open shelves, which have an assigned uncertainty of 50–100% because of the very sparse data coverage. Flux values and their uncertainties were derived according to the methods in Box 1.

degree than those of shelf waters². In addition to explaining the shelf switching from a CO₂ source to a CO₂ sink over the industrial age, this alternative mechanism would also allow for only a small or no change in shelf organic carbon export, whereas shelf DIC export would be greatly enhanced from 0.3 to 0.7 Pg C yr^{-1} between pre- and post-industrial times (Fig. 4a, b). Observations in southeastern US shelf waters⁸⁷ support the predicted strong present-day DIC export across the coastal

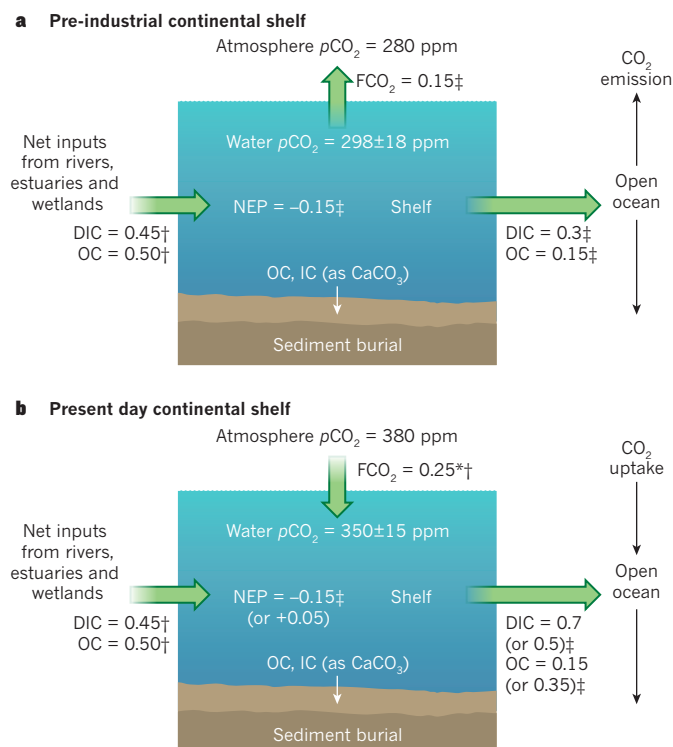


Figure 4 | $p\text{CO}_2$ levels, net ecosystem production and organic and inorganic carbon fluxes in pre-industrial and current continental shelves. According to this model, **a**, the entire pre-industrial continental shelf was a source of CO_2 to the atmosphere, but **b**, it became a CO_2 sink at some point in the latter twentieth century owing to increased atmospheric $p\text{CO}_2$. We further suggest that together with the current known CO_2 uptake from the atmosphere, increased shelf export of dissolved inorganic carbon (DIC) would lead to an increased DIC inventory in the open ocean. Organic carbon, OC; net ecosystem production, NEP; CO_2 flux, FCO_2 . All carbon fluxes including NEP have units of Pg C yr^{-1} . Flux estimates were derived as described in Box 1. We assign an uncertainty of 100% for the pre-industrial air–sea CO_2 flux, larger than the current CO_2 flux uncertainty of 50–75%. *95% certainty that the estimate is within 50% of the reported value; †95% certainty that the estimate is within 100% of the reported value; ‡uncertainty greater than 100% (Box 1). The sea surface average $p\text{CO}_2$ value and its associated current or pre-industrial uncertainties were back-calculated from the assigned air–sea CO_2 flux, known atmospheric $p\text{CO}_2$ value, shelf area and average gas-exchange parameter², as described in Box 1.

open-ocean boundary. Regardless of the specific mechanism, the present-day combined organic and inorganic carbon flux across the shelf break ($0.85 \text{ Pg C yr}^{-1}$) is significantly smaller than the combined inputs (1.3 Pg C yr^{-1}) from rivers, estuaries or wetlands ($0.95 \text{ Pg C yr}^{-1}$) and the atmosphere ($0.25 \text{ Pg C yr}^{-1}$) (Fig. 4b), making the modern coastal ocean, as a whole, an important zone of global carbon sequestration ($0.35 \text{ Pg C yr}^{-1}$) (ref. 3).

The differences in the export fluxes of organic and inorganic carbon by these two proposed mechanisms are great, but subject to large uncertainties (Box 1). In addition to the associated uncertainties in the air–sea CO_2 flux owing to limited data coverage, we also note that to focus on the potential importance of changing CO_2 levels and nutrient enrichment, we have assumed that both organic and inorganic carbon export to shelves and burial in shelf sediments have remained constant (Fig. 4a, b). If instead these export and burial fluxes have increased over the industrial period, then lateral export fluxes could be significantly smaller than those suggested here (see ref. 3 for details).

Development of agricultural soils, deforestation, sewage inputs, enhanced weathering of continental surfaces (Fig. 1a) and loss of coastal wetlands (Fig. 1b) have probably caused changes in both estuarine and continental shelf carbon fluxes over the past century³. Increasing ocean

acidification will also lead to changes in the magnitudes of atmospheric CO_2 exchange and lateral DIC flux by affecting CO_2 and calcium carbonate precipitation and dissolution^{88,89}. The interactions between multiple coastal ocean stressors, such as acidification and eutrophication and the associated hypoxia, may add another level of complexity to predicting future changes in organic and inorganic carbon and CO_2 fluxes in coastal oceans^{2,80}.

Charting the course ahead

Our qualitative and quantitative understanding of organic and inorganic carbon fluxes in the coastal ocean is important for achieving closure on the oceanic and global carbon budgets. Several significant conclusions and questions arise from our analysis of the coastal carbon cycle and the effects of human perturbations that can help to guide future research. For example, only recently has it become clear that the direction and magnitude of estuarine and continental shelf CO_2 exchange with the atmosphere is highly dependent on the terrestrial organic carbon budget and nutrient supplies to the coastal ocean^{2,33} (Figs 1, 2 and Box 3). In consequence, the integration of coastal and global carbon cycles is still in its early stages, and additional efforts are required to fully merge component subsystems such as tidal marshes and mangroves with these budgets^{7,49} (Figs 1b, 3b). Direct measures of wetland carbon exchanges with the atmosphere and export to estuaries, for example using atmospheric eddy covariance techniques, coupled with high temporal resolution export monitoring, will greatly improve this integration. This is particularly important in view of the potential effects of accelerating sea-level rise and global warming on the carbon supply of these ecosystems to the coastal ocean.

We also currently lack the necessary measurements and data to fully categorize the high degree of heterogeneity of coastal systems that would allow extrapolation of our understanding of carbon dynamics at local scales to the global scale. The very large uncertainty associated with estimates of air–water CO_2 fluxes in coastal waters and wetlands (about 50% or $\pm 0.2 \text{ Pg C yr}^{-1}$; Fig. 2) further impedes satisfactory assessment of the overall CO_2 exchange from land and ocean to the atmosphere. The uncertainty in present-day air–surface CO_2 flux estimates in coastal systems must also be reduced before meaningful predictions of the effects of climate change on future fluxes can be made. To resolve these issues, long-term observations and field studies targeted at systems that capture the range of heterogeneity of coastal systems are needed. New technologies, including recent advances in characterizing the sources, ages and reactivity of less stable modern organic carbon (for example, from phytoplankton and modern vascular plants)^{90–93} and new CO_2 , pH and DOC sensors should help to resolve the fate of organic and inorganic carbon in future coastal carbon budgets. Sensitivity analyses of coupled hydrodynamic–biogeochemical model parameters will highlight where additional research focus is needed to further reduce uncertainty in model predictions. These next steps will improve our prediction of coastal carbon dynamics for the full range of estuarine, wetland, and continental shelf subsystems and our understanding of these subsystems' sensitivity to human activities at the larger scales of Earth system models³ (Box 2).

There is growing consensus among Earth system scientists that there are unequivocal anthropogenic signals that can be measured in many of the carbon pools and fluxes in all the major Earth subsystems. Although rivers and their associated watersheds have some of the longest-running evidence of human impacts⁷⁴, we are now recognizing the propagation of these impacts to estuaries and their associated wetlands, as well as to continental shelves. Thus, there is a crucial need for design and implementation of field and modelling activities for carbon-cycle research on coupled terrestrial–ocean systems that bridge, or even eliminate, conventionally defined subsystem boundaries. Nitrogen enrichment of continental shelves from watersheds represents one example of impact propagation across boundaries. Studies designed to examine the mechanisms and time frames of this propagation are needed to assess how future human activities and climate changes will amplify or dampen both the spatial extent of the propagation and its impact.

If the shelf has indeed shifted from an overall pre-industrial CO_2 source

to a present-day and future CO₂ sink, does this also indicate a shift in shelf ecosystem metabolism from a net heterotrophic to a net autotrophic state over the past few decades⁶⁹? This possibility is supported by the massive global increase in nutrient inputs from most rivers to the coastal ocean owing to agricultural fertilizers and other human-related sources on land. Enhanced nutrient inputs have undoubtedly shifted local trophic states both spatially and temporally, and may have even reversed some shelf ecosystems from net heterotrophy to net autotrophy. However, we suggest that coastal systems globally (excluding tidal wetlands) are overall heterotrophic, respiring more organic carbon than they synthesize. The much greater amounts of organic carbon (supporting heterotrophy) relative to inorganic nitrogen (supporting autotrophy) in river waters that discharge into coastal systems (Box 3) support this contention. In addition, losses of inorganic nitrogen (80% or more) by denitrification in coastal systems⁹⁴ further limit autotrophy.

To reconcile the present simultaneous net heterotrophic state and uptake of CO₂, we suggest that continental shelves might still be a CO₂ source to the atmosphere if current atmospheric p_{CO_2} was lower (for example, below 350 p.p.m., the likely present-day average coastal water p_{CO_2} ; Fig. 4b). The reason p_{CO_2} is not higher in today's shelf waters may result from the short residence times (generally less than a few months, and shorter than the air–sea CO₂ equilibration time scale) of their water masses, preventing full accumulation of anthropogenic CO₂ in these waters² (Fig. 4a, b). If p_{CO_2} levels continue to increase faster in the atmosphere than in shelf waters, then the coastal ocean will take up more CO₂ from the atmosphere and export more DIC to the open ocean. This postulation should be examined through long-term time series measurements of carbon fluxes (for example, moored arrays at fixed shelf stations) such as those that are common in open ocean and terrestrial monitoring, and decadal cross-shelf transects in representative coastal systems. If coastal ocean CO₂ uptake and lateral DIC export become increasingly important relative to a stabilizing or decreasing open ocean CO₂ sink⁷¹, this mechanism must be incorporated in future models of ocean–atmosphere CO₂ exchange and ocean acidification. ■

Received 29 June; accepted 28 October 2013.

- Cole, J. J. *et al.* Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10**, 172–185 (2007). **This paper presents a new model of the quantitatively significant roles of inland waters, including streams and rivers and estuaries, in transporting and burying terrestrial carbon, degrading reactive organic carbon, and CO₂ emissions.**
- Cai, W.-J. Estuarine and coastal ocean carbon paradox: CO₂ sinks or sites of terrestrial carbon incineration? *Annu. Rev. Mar. Sci.* **3**, 123–145 (2011).
- Regnier, P. *et al.* Anthropogenic perturbation of the carbon fluxes from land to ocean. *Nature Geosci.* **6**, 597–607 (2013).
- IPCC. *Climate Change 2013. The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) (Cambridge Univ. Press, in the press).
- Aufdenkampe, A. K. *et al.* Riverine coupling of biogeochemical cycles between land, oceans and atmosphere. *Front. Ecol. Environ.* **9**, 53–60 (2011).
- Hopkinson, C. S. *et al.* Terrestrial inputs of organic matter to coastal ecosystems: an intercomparison of chemical characteristics and bioavailability. *Biogeochemistry* **43**, 211–234 (1998).
- Testa, J. M., Kemp, W. M., Hopkinson, C. S. & Smith, S. V. In *Estuarine Ecology* 2nd edn (eds Day, J. W., Crump, B. C., Kemp, W. M. & Yanez-Arancibia, Y.) 381–416 (Wiley, 2013).
- Hofmann, E. E. *et al.* Modeling the dynamics of continental shelf carbon. *Annu. Rev. Mar. Sci.* **3**, 93–122 (2011).
- Mayorga, E. *et al.* Global nutrient export from WaterSheds 2 (NEWS 2): model development and implementation. *Environ. Model. Softw.* **25**, 837–853 (2010).
- Dai, M. H., Yin, Z. Q., Meng, F. F., Liu, Q. & Cai, W. J. Spatial distribution of riverine DOC inputs to the ocean: an updated global synthesis. *Curr. Op. Environ. Sustain.* **4**, 170–178 (2012).
- Syvitski, J. P. M., Vorosmarty, C. J., Kettner, A. J. & Green, P. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. *Science* **308**, 376–380 (2005).
- Hartmann, J., Jansen, N., Durr, H. H., Kempe, S. & Kohler, P. Global CO₂-consumption by chemical weathering: what is the contribution of highly active weathering regions? *Global Planet. Change* **69**, 185–194 (2009).
- Meybeck, M. Carbon, nitrogen and phosphorus transport by world rivers. *Am. J. Sci.* **282**, 401–450 (1982). **The first comprehensive summary of terrestrial export of organic and inorganic C, N and P from mostly natural watersheds of the world.**
- Raymond, P. A. *et al.* Flux and age of dissolved organic carbon exported to the Arctic Ocean: a carbon isotopic study of the five largest arctic rivers. *Glob. Biogeochem. Cycles* **21**, GB4011 (2007).
- Beaulieu, E., Godderis, Y., Donnadieu, Y., Labat, D. & Roelandt, C. High sensitivity of the continental-weathering carbon dioxide sink to future climate change. *Nature Clim. Change* **2**, 346–349 (2012).
- Laudon, H. *et al.* Cross-regional prediction of long-term trajectory of stream water DOC response to climate change. *Geophys. Res. Lett.* **39**, L18404 (2012).
- Hilton, R. G. *et al.* Tropical-cyclone-driven erosion of the terrestrial biosphere from mountains. *Nature Geosci.* **1**, 759–762 (2008).
- Yoon, B. & Raymond, P. A. Dissolved organic matter export from a forested watershed during Hurricane Irene. *Geophys. Res. Lett.* **39**, L18402 (2012).
- Raymond, P. A. & Oh, N. H. An empirical study of climatic controls on riverine C export from three major U.S. watersheds. *Glob. Biogeochem. Cycles* **21**, GB2022 (2007).
- Godsey, S. E., Kirchner, J. W. & Clow, D. W. Concentration-discharge relationships reflect chemostatic characteristics of US catchments. *Hydrol. Processes* **23**, 1844–1864 (2009).
- Bianchi, T. S. *et al.* Enhanced transfer of terrestrially derived carbon to the atmosphere in a flooding event. *Geophys. Res. Lett.* **40**, 116–122 (2013).
- Laruelle, G. G. *et al.* Global multi-scale segmentation of continental and coastal waters from the watersheds to the continental margins. *Hydrol. Earth Syst. Sci.* **17**, 2029–2051 (2013).
- Bauer, J. E. & Bianchi, T. S. in *Treatise on Estuarine and Coastal Science*, Vol. 5 (eds Wolanski, E. & McLusky, D. S.) 7–67 (Academic, 2011).
- Bianchi, T. S. & Bauer, J. E. 2011. in *Treatise on Estuarine and Coastal Science*, Vol. 5 (eds Wolanski, E. & McLusky, D. S.) 69–117 (Academic, 2011).
- Raymond, P. A. & Bauer, J. E. Use of ¹⁴C and ¹³C natural abundances as a tool for evaluating freshwater, estuarine and coastal organic matter sources and cycling. *Org. Geochem.* **32**, 469–485 (2001).
- Raymond, P. A. & Hopkinson, C. S. J. Ecosystem modulation of dissolved carbon age in a temperate marsh-dominated estuary. *Ecosystems* **6**, 694–705 (2003).
- Keil, R. G., Mayer, L. M., Quay, P. E., Richey, J. E. & Hedges, J. I. Loss of organic matter from riverine particles in deltas. *Geochim. Cosmochim. Acta* **61**, 1507–1511 (1997).
- Mayer, L. M., Schick, L. L., Skorko, K. & Boss, E. Photodissolution of particulate organic matter from sediments. *Limnol. Oceanogr.* **51**, 1064–1071 (2006).
- Moran, M. A., Sheldon, W. M. & Zepp, R. G. Carbon loss and optical property changes during long-term photochemical and biological degradation of estuarine dissolved organic matter. *Limnol. Oceanogr.* **45**, 1254–1264 (2000).
- Smith, E. M. & Benner, R. Photochemical transformations of riverine dissolved organic matter: effects on estuarine bacterial metabolism and nutrient demand. *Aquat. Microb. Ecol.* **40**, 37–50 (2005).
- Vanderborght, J. P., Wollast, R., Loijens, M. & Regnier, P. Application of a transport-reaction model to the estimation of biogas fluxes in the Scheldt estuary. *Biogeochemistry* **59**, 207–237 (2002).
- Sholkovitz, E. R. Flocculation of dissolved organic and inorganic matter during the mixing of river water and seawater. *Geochim. Cosmochim. Acta* **40**, 831–845 (1976).
- Borges, A. V. & Abril, G. in *Treatise on Estuarine and Coastal Science*, Vol. 5 (eds Wolanski, E. & McLusky, D. S.) 119–161 (Academic, 2011). **This article is a synthesis of the state of our knowledge of the inorganic carbon cycle in coastal waters.**
- Mantoura, R. F. C. & Woodward, E. Conservative behavior of riverine dissolved organic matter in the Severn estuary. *Geochim. Cosmochim. Acta* **47**, 1293–1309 (1983). **This is the first study to quantitatively assess riverine DOC supply and fate in estuaries using salinity as a conservative tracer and to assess the reactivity of this DOC in ocean carbon budgets.**
- Hopkinson, C. S. & Vallino, J. J. The relationship between man's activities in watersheds and estuaries: a model of runoff effects on patterns of ecosystem metabolism. *Estuaries* **18**, 598–621 (1995).
- Howarth, R. *et al.* Coupled biogeochemical cycles: eutrophication and hypoxia in temperate estuaries and coastal marine ecosystems. *Front. Ecol. Environ.* **9**, 18–26 (2011).
- Hofmann, A. F., Soetaert, K. & Middelburg, J. J. Present nitrogen and carbon dynamics in the Scheldt estuary using a novel 1-D model. *Biogeosciences* **5**, 981–1006 (2008).
- Frankignoulle, M. *et al.* Carbon dioxide emission from European estuaries. *Science* **282**, 434–436 (1998). **The first regional synthesis of CO₂ emission fluxes from estuaries, demonstrating their importance to regional carbon budgets.**
- Borges, A. V., Delille, B. & Frankignoulle, M. Budgeting sinks and sources of CO₂ in the coastal ocean: diversity of ecosystems counts. *Geophys. Res. Lett.* **32**, L14601 (2005).
- Laruelle, G. G. & Dürr, H. H. Slomp, C. P. & Borges, A. V. Evaluation of sinks and sources of CO₂ in the global coastal ocean using a spatially-explicit typology of estuaries and continental shelves. *Geophys. Res. Lett.* **37**, L15607 (2010).
- Schlesinger, W. H. & Bernhardt, E. S. *Biogeochemistry, An Analysis of Global Change* 3rd edn (Academic, 2013).
- Cai, W.-J. & Wang, Y. The chemistry, fluxes, and sources of carbon dioxide in the estuarine waters of the Satilla and Altamaha Rivers, Georgia. *Limnol. Oceanogr.* **43**, 657–668 (1998).
- Raymond, P. A., Caraco, N. F. & Cole, J. J. Carbon dioxide concentration and atmospheric flux in the Hudson River. *Estuaries* **20**, 381–390 (1997).
- Cai, W.-J., Wang, Y. C., Krest, J. & Moore, W. S. The geochemistry of dissolved inorganic carbon in a surficial groundwater aquifer in North Inlet, South Carolina, and the carbon fluxes to the coastal ocean. *Geochim. Cosmochim. Acta* **67**, 631–639 (2003).

45. Cai, W.-J. Riverine inorganic carbon flux and rate of biological uptake in the Mississippi River plume. *Geophys. Res. Lett.* **30**, 1032 (2003).
46. Zhai, W., Dai, M. & Guo, X. Carbonate system and CO₂ degassing fluxes in the inner estuary of Changjiang (Yangtze) River, China. *Mar. Chem.* **107**, 342–356 (2007).
47. Dittmar, T., Hertkorn, N., Kattner, G. & Lara, R. Mangroves, a major source of dissolved organic carbon to the oceans. *Glob. Biogeochem. Cycles* **20**, GB1012 (2006).
48. Mcleod, E. *et al.* A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO₂. *Front. Ecol. Environ.* **9**, 552–560 (2011).
49. Hopkinson, C. S., Cai, W. J. & Hu, X. Carbon sequestration in wetland dominated coastal systems - a global sink of rapidly diminishing magnitude. *Curr. Op. Environ. Sustain.* **4**, 1–9 (2012).
50. Mackenzie, F. T., Andersson, A. J., Lerman, A. & Ver, L. M. in *The Sea Vol. 13* (eds Robinson, A. R. & Brink, K. H.) 193–225 (Harvard Univ. Press, 2005).
- This is a comprehensive historical description of carbon cycling processes and fluxes through Earth's past, present and future.**
51. Jahnke, R. in *Carbon and Nutrient Fluxes in Continental Margins* (eds Liu, K. K., Atkinson, L., Quinones, R. & Talaure-McManus, L.) 597–615 (Springer, 2010).
52. Liu, K. K., Atkinson, L., Quinones, R. & Talaure-McManus, L. *Carbon and Nutrient Fluxes in Continental Margins* (Springer, 2010).
53. Azetsu-Scott, K. & Passow, U. Ascending marine particles: significance of transparent exopolymer particles (TEP) in the upper ocean. *Limnol. Oceanogr.* **49**, 741–748 (2004).
54. Blair, N. E. & Aller, R. C. The fate of terrestrial organic carbon in the marine environment. *Annu. Rev. Mar. Sci.* **4**, 401–423 (2012).
55. Bauer, J. E., Williams, P. M. & Druffel, E. R. M. ¹⁴C activity of dissolved organic carbon fractions in the central North Pacific and Sargasso Sea. *Nature* **357**, 667–670 (1992).
56. Stubbins, A. *et al.* Illuminated darkness: molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. *Limnol. Oceanogr.* **55**, 1467–1477 (2010).
57. Hertkorn, N. *et al.* Characterization of a major refractory component of marine dissolved organic matter. *Geochim. Cosmochim. Acta* **70**, 2990–3010 (2006).
58. Hernes, P. J. & Benner, R. Photochemical and microbial degradation of dissolved lignin phenols: implications for the fate of terrigenous dissolved organic matter in marine environments. *J. Geophys. Res. Oceans* **108**, 3291 (2003).
59. Bianchi, T. S. The role of terrestrially derived organic carbon in the coastal ocean: A changing paradigm and the priming effect. *Proc. Natl Acad. Sci. USA* **108**, 19473–19481 (2011).
60. Tsunogai, S., Watanabe, S. & Sato, T. T. Is there a “continental shelf pump” for the absorption of atmospheric CO₂? *Tellus* **51B**, 701–712 (1999).
61. Cai, W.-J., Dai, M. & Wang, Y. Air–sea exchange of carbon dioxide in ocean margins: A province-based synthesis. *Geophys. Res. Lett.* **33**, L12603 (2006).
62. Jiang, L.-Q., Cai, W.-J., Wanninkhof, R., Wang, Y. & Lüger, H. Air–sea CO₂ fluxes on the U.S. South Atlantic Bight: spatial and seasonal variability. *J. Geophys. Res.* **113**, C07019 (2008).
63. Walsh, J. J. Importance of continental margins in the marine biogeochemical cycling of carbon and nitrogen. *Nature* **350**, 53–55 (1991).
- This is an influential paper summarizing a major US-funded effort, concluding that shelf-break mixing provides a mechanism for CO₂ loss from ocean margins.**
64. Huang, W.-J., Cai, W.-J., Castelao, R. M., Wang, Y. & Lohrenz, S. E. Effects of a wind-driven cross-shelf large river plume on biological production and CO₂ uptake in the Gulf of Mexico during spring. *Limnol. Oceanogr.* **58**, 1727–1735 (2013).
65. Hales, B., Takahashi, T. & Bandstra, L. Atmospheric CO₂ uptake by a coastal upwelling system. *Glob. Biogeochem. Cycles* **19**, GB1009 (2005).
66. Chen, C.-T. A., Huang, T.-H., Chen, Y.-C., Bai, Y., He, X., & Kang, Y. Air–sea exchanges of CO₂ in world's coastal seas. *Biogeosciences* **10**, 5041–5105 (2013).
67. Bates, N. R., Moran, S. B., Hansell, D. A. & Mathis, J. T. An increasing CO₂ sink in the Arctic Ocean due to sea-ice loss. *Geophys. Res. Lett.* **33**, L23609 (2006).
68. Smith, S. V. & Hollibaugh, J. T. Coastal metabolism and the oceanic carbon balance. *Rev. Geophys.* **31**, 75–89 (1993).
- One of the first papers to show that coastal oceans are net heterotrophic due to respiration enhanced by exported terrestrial oceanic carbon.**
69. Mackenzie, F. T., Lerman, A. & Andersson, A. J. Past and present of sediment and carbon biogeochemical cycling models. *Biogeosciences* **1**, 11–32 (2004).
70. del Giorgio, P. A. & Williams, P. J. *Respiration in Aquatic Ecosystems* (Oxford Univ. Press, 2005).
71. Le Quéré *et al.* The global carbon budget 1959–2011. *Earth Syst. Sci. Data* **5**, 165–185 (2013).
72. Stallard, R. F. Terrestrial sedimentation and the carbon cycle: Coupling weathering and erosion to carbon burial. *Glob. Biogeochem. Cycles* **12**, 231–257 (1998).
73. Smith, S. V., Renwick, W. H., Buddemeier, R. W. & Crossland, C. J. Budgets of soil erosion and deposition for sediments and sedimentary organic carbon across the conterminous United States. *Glob. Biogeochem. Cycles* **15**, 697–707 (2001).
74. Raymond, P. A., Oh, N. H., Turner, R. E. & Broussard, W. Anthropogenically enhanced fluxes of water and carbon from the Mississippi River. *Nature* **451**, 449–452 (2008).
75. Wilson, H. F. & Xenopoulos, M. A. Effects of agricultural land use on the composition of fluvial dissolved organic matter. *Nature Geosci.* **2**, 37–41 (2009).
76. Raymond, P. A. *et al.* Controls on the variability of organic matter and dissolved inorganic carbon ages in northeast US rivers. *Mar. Chem.* **92**, 353–366 (2004).
77. Moore, S. *et al.* Deep instability of deforested tropical peatlands revealed by fluvial organic carbon fluxes. *Nature* **493**, 660–663 (2013).
78. Monteith, D. T. *et al.* Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. *Nature* **450**, 537–540 (2007).
79. Calmels, D., Gaillardet, J., Brenot, A. & France-Lanord, C. Sustained sulfide oxidation by physical erosion processes in the Mackenzie River basin: climatic perspectives. *Geology* **35**, 1003–1006 (2007).
80. Cai, W.-J. *et al.* Acidification of subsurface coastal waters enhanced by eutrophication. *Nature Geosci.* **4**, 766–770 (2011).
81. Doney, S. C. The growing human footprint on coastal and open-ocean biogeochemistry. *Science* **328**, 1512–1516 (2010).
- A paper describing the close coupling between land and ocean systems and the importance of anthropogenic impacts on coastal and open ocean systems.**
82. Dai, M. *et al.* Oxygen depletion in the upper reach of the Pearl River estuary during a very drought winter. *Mar. Chem.* **102**, 159–169 (2006).
83. Scavia, D. *et al.* Climate change impacts on U.S. coastal and marine ecosystems. *Estuaries* **25**, 149–164 (2002).
84. Stedmon, C. A., Amon, R. M. W., Rinehart, A. J. & Walker, S. A. The supply and characteristics of colored dissolved organic matter (CDOM) in the Arctic Ocean: pan Arctic trends and differences. *Mar. Chem.* **124**, 108–118 (2011).
85. Diaz, R. J. & Rosenberg, R. Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926–929 (2008).
86. Duarte, C. M., Middelburg, J. & Caraco, N. Major role of marine vegetation on the oceanic carbon cycle. *Biogeosciences* **2**, 1–8 (2005).
- This paper describes a multifaceted approach to quantifying the carbon fixed by estuarine macrophytes and its relevance to organic carbon sequestration in coastal sediments and heterotrophy in adjacent coastal systems.**
87. Cai, W.-J., Wang, Z. H. A. & Wang, Y. C. The role of marsh-dominated heterotrophic continental margins in transport of CO₂ between the atmosphere, the land-sea interface and the ocean. *Geophys. Res. Lett.* **30**, 1849 (2003).
88. Andersson, A. J., Mackenzie, F. T. & Lerman, A. Coastal ocean and carbonate systems in the high CO₂ world of the Anthropocene. *Am. J. Sci.* **305**, 875–918 (2005).
89. Krumins, V., Gehlen, M., Arndt, S., Van Cappellen, P. & Regnier, P. Dissolved inorganic carbon and alkalinity fluxes from coastal marine sediments: model estimates for different shelf environments and sensitivity to global change. *Biogeosciences* **10**, 371–398 (2013).
90. Bianchi, T. S. & Canuel, E. A. *Chemical Biomarkers in Aquatic Ecosystems* (Princeton Univ. Press, 2011).
91. Ingalls, A. E. & Pearson, A. Ten years of compound-specific radiocarbon analysis. *Oceanography* **18**, 18–31 (2005).
92. Mopper, K., Stubbins, A., Ritchie, J. D., Bialk, H. M. & Hatcher, P. G. Advanced instrumental approaches for characterization of marine dissolved organic matter: extraction techniques, mass spectrometry, and nuclear magnetic resonance spectroscopy. *Chem. Rev.* **107**, 419–442 (2007).
93. Kujawinski, E. B. The impact of microbial metabolism on marine dissolved organic matter. *Annu. Rev. Mar. Sci.* **3**, 567–599 (2011).
94. Seitzinger, S. P. & Giblin, A. E. Estimating denitrification in North Atlantic continental shelf sediments. *Biogeochemistry* **35**, 235–260 (1996).
95. Collins, W. J. *et al.* Development and evaluation of an Earth-system model–HadGEM2. *Geosci. Model Dev.* **4**, 1051–1075 (2011).
96. da Cunha, L. C., Buitenhuis, E. T., Le Quéré, C., Giraud, X. & Ludwig, W. Potential impact of changes in river nutrient supply on global ocean biogeochemistry. *Glob. Biogeochem. Cycles* **21**, GB4007 (2007).
97. Bernard, C., Dürr, H., Heinze, C., Segsneider, J. & Maier-Reimer, E. Contribution of riverine nutrients to the silicon biogeochemistry of the global ocean – a model study. *Biogeosciences* **8**, 551–564 (2011).
98. Lachkar, Z. & Gruber, N. What controls biological productivity in coastal upwelling systems? Insights from a comparative modeling study. *Biogeosciences* **8**, 5617–5652 (2011).
99. Chen, R. F., Fry, B., Hopkinson, C. S., Repeta, D. J. & Peltzer, E. T. Dissolved organic carbon on Georges Bank. *Cont. Shelf Res.* **16**, 409–420 (1996).
100. Hopkinson, C. S. & Vallino, J. J. Efficient export of carbon to the deep ocean through dissolved organic carbon. *Nature* **433**, 142–145 (2005).

Acknowledgements This work was supported in part by the National Science Foundation's Chemical and Biological Oceanography, Integrated Carbon Cycle Research, Arctic Natural Sciences, Long-Term Ecological Research, and Ecosystem Ecology programs; NASA Interdisciplinary Research in Earth Science program NOAA; Georgia Sea Grant; the European Union's Seventh Framework Program project GEOCARBON; and the government of the Brussels-Capital Region. We acknowledge our late friend and colleague Y. Wang, whose contributions to coastal carbon cycle research and CO₂ measurement technology have significantly advanced the field. We also thank A. Grotoli for comments and discussion on an earlier version of this manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/fir3il. Correspondence should be addressed to J.E.B. (bauer.362@osu.edu).

Offshore fresh groundwater reserves as a global phenomenon

Vincent E.A. Post^{1,2}, Jacobus Groen^{3,4}, Henk Kooi³, Mark Person⁵, Shemin Ge⁶ & W. Mike Edmunds⁷

The flow of terrestrial groundwater to the sea is an important natural component of the hydrological cycle. This process, however, does not explain the large volumes of low-salinity groundwater that are found below continental shelves. There is mounting evidence for the global occurrence of offshore fresh and brackish groundwater reserves. The potential use of these non-renewable reserves as a freshwater resource provides a clear incentive for future research. But the scope for continental shelf hydrogeology is broader and we envisage that it can contribute to the advancement of other scientific disciplines, in particular sedimentology and marine geochemistry.

Continental shelves are the submerged fringes of the continents and harbour important aquifers beneath the sea floor. Because the shelves are at present covered by sea water, hydrogeology — a scientific discipline with an almost exclusive focus on fresh terrestrial groundwater resources — has conventionally paid little attention to them¹. But on a geological timescale, the realm of the terrestrial hydrological cycle has been expanding and contracting as coastlines migrated² with the falling and rising of global sea levels³. The exposure of the shelves reached its most recent maximum during the Last Glacial Maximum, from 26,500 to about 19,000 years ago⁴. Shelf areas that were exposed during sea-level low-stands were covered by freshwater lake and river systems^{5,6}, and were subject to the infiltration of atmospheric precipitation² (also called meteoric water) and glacial meltwater. This led to extensive emplacement and circulation of fresh groundwater.

Groundwater systems are slow to adapt to the reconfiguration of the hydrological conditions at Earth's surface^{7–9}, and therefore remnants of meteoric groundwater are likely to be found offshore. Now that it is becoming clear that anthropogenic and natural changes in continental water storage affect global sea level^{10,11}, and that the sequestration of fresh water below continental shelves contributed to the increase of ocean salinity during glacial periods¹², an appraisal of offshore groundwater as an element in global environmental change is warranted. Moreover, because continental shelf aquifers underlie areas that are in a continuous state of transition in response to global climate and sea level, offshore groundwater could hold important clues to the natural variability of the hydrological cycle over thousands of years, or even longer.

In this Review, we discuss overwhelming evidence that vast meteoric groundwater reserves (VMGRs) below the sea floor are a common global phenomenon and review the recent advances in our understanding of the key mechanisms that favour the emplacement, as well as the preservation, of VMGRs. The salinity within VMGRs can range between that of fresh water and that of sea water, and their delineation requires a practical definition. VMGRs are defined in this Review as a groundwater body with a minimum horizontal extent of 10 km, and a minimum concentration of total dissolved solids (TDS) less than 10 g l⁻¹, which is about one-third of the salinity of sea water.

The selection of this salinity threshold is deliberate — it coincides with the upper limit of the salinity range used for the definition of brackish water in the area of water desalination¹³. Brackish water is increasingly

seen as a resource for water supply^{14,15} because the energy needs of reverse osmosis¹⁶, and therefore costs of desalination, are decreasing. The widespread confirmation of the scale of offshore fresh and brackish groundwater reserves therefore provides opportunities for the relief of water scarcity in densely populated coastal regions. Offshore groundwater abstraction can help to mitigate the adverse effects of onshore pumping, such as land subsidence^{17,18} and seawater intrusion^{19,20}. This provides another important impetus to shift the boundaries of hydrogeology into the offshore domain.

Limits of modern coastal groundwater systems

It has long been known that the coastline does not form a boundary for coastal groundwater systems¹⁴. Sea water can intrude inland^{19,20}, and land-derived fresh groundwater may discharge through the sea floor through a process known as submarine groundwater discharge^{21,22} (SGD). Myriad studies have highlighted the ubiquitous occurrence of SGD (for example, see ref. 23), but most SGD studies have focused on the near-shore environment^{22,24}, and we still need to understand the groundwater conditions and processes beneath the continental shelves further offshore¹⁴.

Hydrological modelling studies^{25,26} have shown that SGD can extend far beyond the coastline in aquifers that are separated from the sea by a confining layer of low permeability. Groundwater from the submarine aquifer discharges slowly by upward flow through the confining layer across extensive areas²². The Indian River Bay in Delaware²⁷ is a well-characterized example, where fresh water occurs up to 1 km offshore in a confined sandy aquifer. For carbonate aquifers (made up of limestone or dolomite) with dissolution-formed flow conduits, discharge in the form of submarine freshwater springs is a well-known phenomenon^{22,23}.

In the carbonate aquifer system along the eastern seaboard of Florida (Fig. 1), fresh water found in boreholes up to 100 km from the coast²⁸ (Fig. 2) has also been linked to high water table conditions that existed at the northern seaboard of Florida before the time of major groundwater exploitation²⁵, suggesting that SGD extends over a distance of 100 km or more. Observed pressures of sub-seafloor fresh waters — fresh water can rise up to 9 m above sea level in boreholes — are consistent²⁸ with this interpretation. However, the buoyancy of a large freshwater body surrounded by saline groundwater can also account for these observations. In other words, not all fresh groundwater below the sea floor is necessarily related to active SGD systems that originate onshore. This seems a likely

¹School of the Environment, Flinders University, PO Box 2100, Adelaide SA 5001, Australia. ²National Centre for Groundwater Research and Training, GPO Box 2100, Adelaide SA 5001, Australia.

³VU University Amsterdam, Faculty of Earth and Life Sciences, Critical Zone Hydrology Group, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands. ⁴Acacia Water, Jan van Beaumontstraat 1, 2805 RN, Gouda, the Netherlands. ⁵New Mexico Tech, Department of Earth & Environmental Science, 801 Leroy Place, Socorro, NM 87801, USA. ⁶University of Colorado, Department of Geological Sciences, Boulder, Colorado 80309, USA. ⁷University of Oxford, School of Geography and the Environment, South Parks Road, Oxford OX1 3QY, UK.

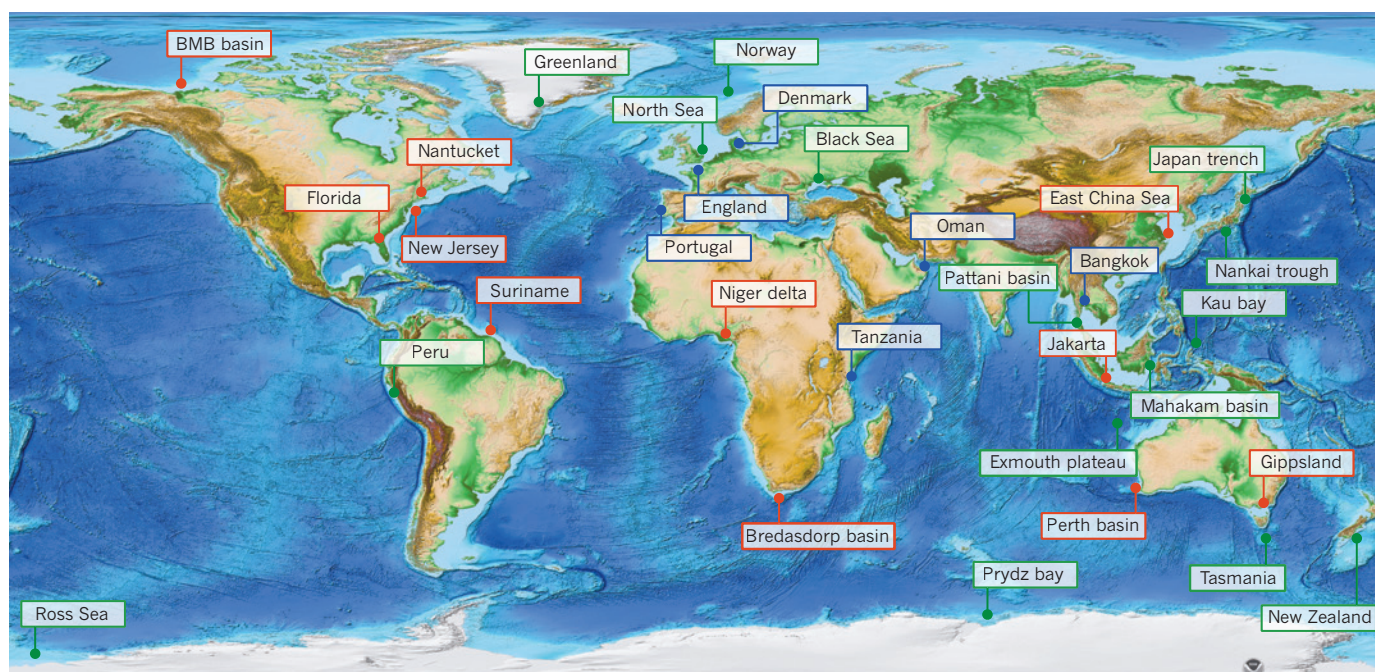


Figure 1 | World map of topography and bathymetry showing known occurrences of fresh and brackish offshore groundwater. Bathymetry from ref. 98. The occurrence of vast meteoric groundwater reserves (VMGRs) proven by direct observational data are shown in red. Offshore groundwater that is not necessarily fresh, but for which a freshwater mixing component has been inferred on the basis of pore-water composition is shown in green.

option because low-salinity groundwater has also been encountered in offshore areas in which active SGD is not possible (as indicated by an onshore water table that is too low to provide enough driving force²⁶ or by the absence of a hydraulic connection with an onshore recharge area²⁹). Such low-salinity water occurrences must therefore be relics of flow systems that sequestered fresh water under different climate, morphology and sea-level conditions, and are referred to as palaeo-groundwater.

Global occurrences of offshore VMGRs

The best-documented example of an offshore palaeo-groundwater body is the vast occurrence of low-salinity water extending below the continental shelf of New Jersey^{30–32} (Fig. 1). Groundwater with a salinity equal to about a quarter of seawater salinity was found up to 100 km offshore³², and later drilling documented freshwater influences up to 130 km from the New Jersey coast³⁰ (Fig. 2 and Table 1). Salinity and pressure data from a deep borehole²⁹ and geophysical data on Nantucket Island, Massachusetts³³, as well as offshore salinity data³² provide further indications for the extensive occurrence of low-salinity palaeo-water beneath the continental shelf of the north-eastern United States (Fig. 2 and Table 1).

Although the Atlantic seaboard of North America provided the first documented studies of offshore VMGRs, there is now ample evidence that VMGRs are a global phenomenon^{34–41} (Fig. 1 and Table 1). Not all VMGRs seem to be connected to onshore aquifers^{40,41}, but it has been inferred that those that do are wedge-shaped, becoming thinner and more saline with increasing distance offshore^{28,30,32,35–37} (Fig. 2). A conspicuous feature of the Suriname³⁵, New Jersey³⁰, Gippsland, Australia³⁶, and Jakarta³⁷ VMGRs is that the transition from high salinities below the sea floor to low salinities in the wedge is narrower than the transition zone from fresh water to salt water at greater depth (Fig. 2).

Some studies have found that the distribution of low-salinity water within VMGRs is controlled by geological features, such as faults and low-permeability layers (for example, the Perth Basin⁴²) or palaeo-channels (for example, East China Sea⁴³ or Bredasdorp Basin in South Africa⁴¹). These examples of implied structural and stratigraphic controls on VMGRs attest to the fact that salinity distributions of VMGRs can

be complex and that pervasive, wedge-shaped interpretations^{28,30,32,35–37} may be oversimplified. This is borne out by recent borehole data off New Jersey³¹, which revealed a complex geometry of vertically alternating freshwater–saltwater intervals that are difficult to correlate at distances of about 10 km.

At various sites around the world, pore-water profiles in low-permeability layers that start just below the sea floor and show a consistent vertical salinity decrease have been documented (Fig. 1 and Table 1). These are found on continental shelves that were exposed during the last glacial period (in the North Sea⁴³, Peru⁴⁴ and New Zealand⁴⁵), or where there used to be lakes when the sea level was lower (Black Sea⁵ and Kau Bay, Indonesia⁴⁶). At these locations, they are probably indicators of former meteoric water circulation.

Genesis

Modelling of selected cases has demonstrated that the location of the freshwater and saltwater transition zone is further offshore than would be expected on the basis of current sea-level and hydrological boundary conditions^{8,26,29,34,47,48}. On the basis of this, and the overwhelming evidence from the field, the most ubiquitously proposed mechanism to explain the presence of fresh water is that it was emplaced during sea-level low-stands that occurred throughout the Pliocene and Pleistocene epoch³. The lower sea level is generally thought to have resulted in steeper water tables^{2,49–51}, thereby enhancing so-called topography-driven groundwater flow and meteoric recharge that occurs on exposed continental shelves (Fig. 3). This is corroborated by the finding that the volume of offshore fresh water seems to be inversely correlated with present-day sea depth⁴⁸, because shelf bathymetry controls the width of shelf exposure during seawater low-stands.

Favourable factors for freshwater emplacement have been inferred from observed salinity patterns and numerical modelling; these include groundwater flow along permeable faults³⁹ and the existence of distal aquifer outcrops, which allow for lateral groundwater flow rates that are relatively higher than in continental shelf aquifers encased in finer-grained deposits⁴⁸. Some studies have found that a fall in sea level alone

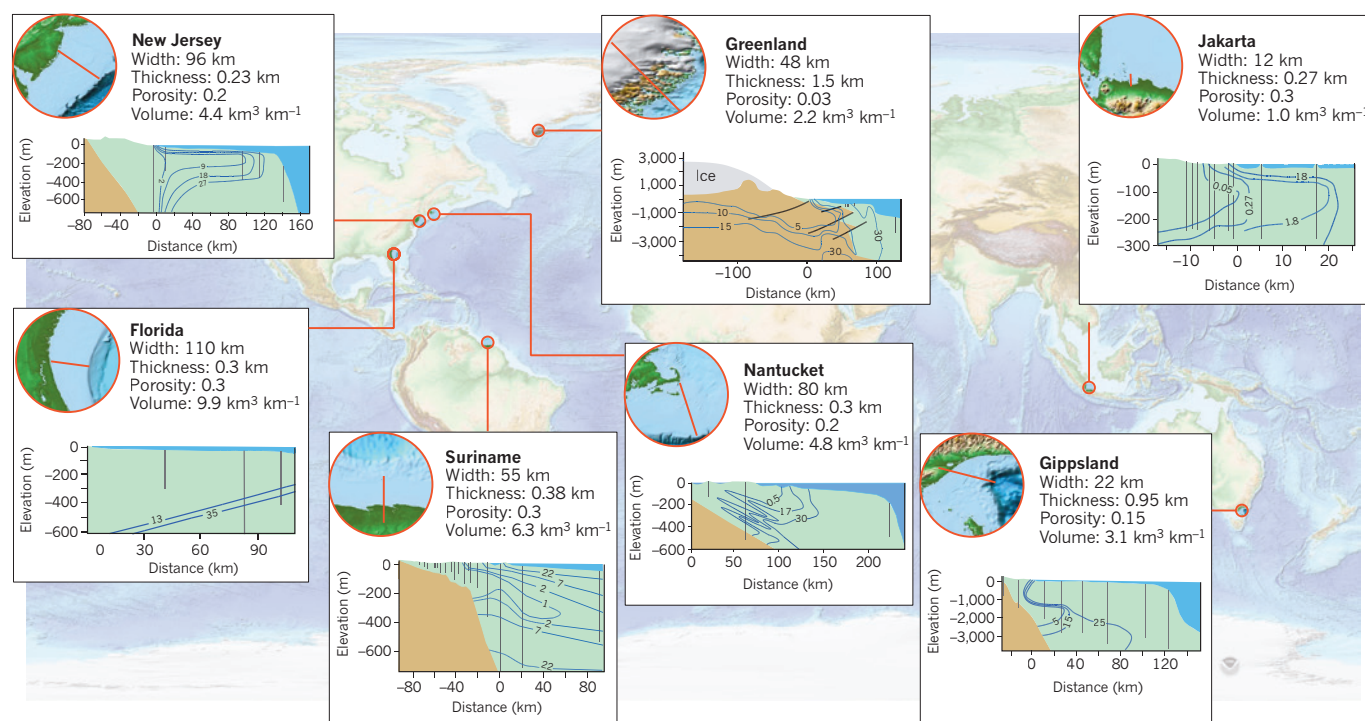


Figure 2 | Global overview of inferred key metrics and cross sections of well-characterised vast meteoric groundwater reserves. Data sourced from refs 28, 30, 32, 35–37, 48, 53. The location of each cross-section is indicated by a red line. In the cross-sections, the blue contour lines indicate total dissolved solid (TDS) concentrations (g l^{-1}); distance (km) and elevation (m) relative to mean sea level are indicated along the horizontal and vertical axis, respectively; vertical grey lines indicate well locations where salinity is inferred from water samples and borehole logs;

crystalline bedrock or low-permeability sedimentary rocks containing salt groundwater are shown in brown; the black, sub-horizontal lines denote faults; undifferentiated continental shelf sediments are in pale green; and sea water is pale blue. Within the Nantucket and Greenland cross sections, salinity contours are based on numerical model results^{48,53} and well data. The inferred widths, lengths and volumes per kilometre of coastline pertain to the groundwater with a TDS concentration less than 10 g l^{-1} (Box 1).

was not sufficient to drive fresh groundwater to depths and the outward regions of the continental shelf where it is found today^{29,34,48}. Incisions by rivers³⁴ (Fig. 3) and the existence of palaeo-valleys⁷ are thought to have provided more localized relief and vigorous topography-driven groundwater flow systems that resulted in deep and extensive flushing of the shelf sediments. This hypothesis is supported by submerged geomorphological features such as spring-derived carbonate mounds² and groundwater-related erosion in submarine canyons⁵², which testify to past groundwater discharge at continental shelves around the world².

At high latitudes, retreating ice sheets probably supplied additional fresh water to the continental shelf environment, a behaviour which has been inferred from numerical modelling^{9,29,53}, but also, for example, from the composition of groundwater 100 km offshore of southeastern Greenland⁵³. The same mechanism could also explain the low pore-water salinity observed at sites off the coast of Antarctica (for example, the Ross Sea⁵⁴ and Prydz Bay⁵⁵). The presence of proglacial lakes³³ has also been suggested to play a part in facilitating the emplacement of fresh water below the continental shelf of New England. Former freshwater lakes or inland seas have also been linked to fresh submarine groundwater in warmer regions, such as the Black Sea⁵ and Indonesia⁴⁶.

The higher than freshwater density of saline groundwater (about 2.5% for sea water) would have impeded the freshening of continental shelf aquifers, because the density limits the depth to which fresh water can flush out saline groundwater. Moreover, it forces fresh water to flow upward along a sloping wedge of saline groundwater, thus reducing the effectiveness of freshwater flow to displace the saline groundwater. Systematic modelling studies of these processes and associated flushing timescales have so far not been published.

Shelf exposure during the last glacial maximum provided an area for terrestrial groundwater recharge that was larger than the area of the present-day land mass by around 10%¹², increasing the potential for

recharge. But the propensity for replenishment by meteoric water strongly depended on local conditions of climate and vegetation. In southern Australia, recharge rates between 10,000–20,000 years ago have been inferred to be higher owing to cooler conditions and the concomitant lower evapotranspiration⁵⁰, whereas in southwestern Europe, the persistence of Atlantic air circulation resulted in continuity of recharge⁵⁶. But recharge was much reduced in northern Africa owing to declining monsoon rains⁵⁶, and in northern Europe because of permafrost conditions^{9,56}.

Distinct pore-water salinity decreases have also been reported for very deep-water sites where the sea floor did not become dry during the Pliocene and Pleistocene^{57–62}. In some cases, for example along convergent plate boundaries (for example, Peru⁵⁸, Nankai Trough⁵⁸ and the Japan Trench)⁶⁰ (Fig. 1 and Table 1), these can be related to water-releasing geochemical processes — such as the dissociation of gas hydrates⁶³, or the dehydration and transformation of hydrous minerals^{58,60}. But in other cases major unresolved issues remain for which such internal water-producing processes cannot account for the observed decrease of pore-water chloride concentrations^{58,60,62,63}, especially where the distance to a land mass is so great that terrestrial groundwater input is highly improbable (for example, Norway⁶¹, and Tasmania⁶² and the Exmouth Plateau⁵⁷ in Australia; Fig. 1 and Table 1). Emplacement during as far back as the Miocene epoch has also been proposed for some occurrences^{39,59}, but it is questionable if preservation of low salinities over such long time frames is tenable⁵⁹.

Preservation

The amount of fresh groundwater sequestered in continental shelves during the last glacial period must have been higher than the volume that is found there today, because part of the fresh water was displaced and salinized by the flooding of the exposed shelves during the Holocene^{12,29,32,48,64,65}, when sea level rose to the present high level. Both

Table 1 | Key metrics of offshore meteoric groundwater occurrences

Location (reference)	Offshore distance (km)	Depth (km)	Minimum TDS (g l ⁻¹)	Observation type	Number of offshore observations	Water depth (m)	Onshore connection
VMGRs							
BMB Basin ³⁹	<100	3–4.5	1.0	WS	>20	<60	Unclear
Bredasdorp Basin ⁴¹	80–120	2–2.5	2	LOG	>30	100–120	No
Florida ^{28,32}	100	0.2–0.6	13	WS	3	<100	Yes
Gippsland ³⁶	70	1–4	5	LOG/WS	6	<100	Yes
Jakarta ³⁷	18	<0.3	0.3	WS	2	<10	Yes
Nantucket ^{29,33}	60	<0.6	0.1	WS	3	<100	Yes
New Jersey ^{30–32}	130	<0.6	1.5	WS	4	<100	Yes
Niger delta ³⁸	40	0.1–2	0.2	LOG	11	<100	Yes
Perth Basin ⁴²	50	1.3–4.0	5	LOG/WS	11	<100	Yes
East China Sea ⁴⁰	60–100	<0.2	1.0	WS	2	10–15	Unclear
Suriname ³⁵	90	<0.6	0.9	LOG	3	<50	Yes
Other offshore groundwater							
Black Sea ⁵	95	<0.03	2	WS	1	350	Unclear
Exmouth Plateau ⁵⁷	225–300	0.85–1	28	WS	2	~1,500	Unclear
Greenland ⁵³	50	<0.25	26	WS	3	400–450	Unclear
Kau Bay ⁴⁶	10	<0.01	27	WS	3	300–450	Unclear
Mahakam Basin ⁹⁷	20	1–3	2	WS	27	<50	Unclear
Nankai Trough ⁵⁸	125	<1.3	29	WS	1	4,675	No
Japan Trench ⁶⁰	115	<1.2	19	WS	1	2,681	No
New Zealand ⁴⁵	42	<0.3	24	WS	1	84	Unclear
North Sea ⁴³	100	<0.006	16	WS	1	36	Unlikely
Norway ⁶¹	300	<0.5	30	WS	2	1,300–1,450	Unlikely
Pattani Basin ⁹⁶	150–200	1.6–2.65	0.3	WS	20	50–70	Unclear
Peru shelf ⁴⁴	20	<0.01	13	WS	1	96	Unclear
Peru shelf ⁵⁹	65	<0.35	20	WS	1	460	Unclear
Peru deep ⁵⁹	200	<0.6	28	WS	3	3,000–5,000	No
Prydz Bay ⁵⁵	60	<0.224	31	WS	1	792	Unclear
Ross Sea ⁵⁴	350	<0.380	27	WS	2	619–633	Unclear
Tasmania ⁶²	70–550	0.2–0.95	28	WS	4	2,147–2,705	No
Onshore indicators							
Denmark ⁷²	NA	0.06–0.3	0.2	WS	NA	NA	NA
England ⁷⁰	NA	0.3–0.4	0.0	WS	NA	NA	NA
Oman ⁷⁶	NA	0.2–0.4	0.2	WS	NA	NA	NA
Portugal ⁷¹	NA	~0.3	0.1	WS	NA	NA	NA
Tanzania ⁷³	NA	0.61	1.3	WS	NA	NA	NA

Depth is relative to the sea floor or land surface for offshore and onshore locations, respectively. Observation type indicates if salinities were inferred from geophysical borehole log data (LOG) or water samples (WS; obtained by either pumping or squeezing of pore water from sediments). When total dissolved solids (TDS) was not reported, it was estimated by multiplying the reported chloride concentration by 1.8 (the TDS/chloride ratio of seawater). Locations are shown in Fig. 1. NA, not applicable.

horizontal, landward migration of the freshwater and saltwater transition zone^{8,47,65} and salinization by different modes of vertical, downward transport of salt from the sea floor^{29,35,64,65} contributed to the reduction of the freshwater volumes. The upper limit of the driving force for landward migration of the freshwater and saltwater transition zone is controlled by the gradient of the continental shelf⁶⁵, and consequently transition zone migration rates are unlikely to have exceeded 10 km per 10 Kyr in high-permeability near-surface (or seafloor) aquifers⁶⁵. Even lower rates are expected to have existed in deeper confined units^{9,29,47,66}. Coastline migration, therefore, outpaced transition zone migration during various stages of the Holocene at most continental margins, carrying sea water on top of fresh groundwater, and causing downward salinization (Fig. 3). This is borne out by several cases in which sea water encroached on clay-rich

or other low-permeability units, and in which downward salinization occurred by molecular diffusion^{5,35,43,46,65} — a slow process in which a time period longer than the Holocene (11 Kyr) would have been required for salinity at a depth of 10 m beneath the sea floor to increase to around one-fifth of seawater salinity. Where the sea flooded more permeable strata, relatively fast downward salinization or seawater intrusion to the base of the seafloor aquifer must have taken place by convective mixing (Fig. 3), whereby dense saltwater plumes sink into the aquifer, displacing fresh water, rising up and discharging through the sea floor^{64,65}.

The importance of seafloor sediments in controlling vertical salinization during periods of sea-level rise is analogous to and exemplified by their role in SGD^{21,22}. The occurrence of fresh groundwater below Indian River Bay, Delaware, for instance, was found to be restricted

to areas where low-permeability sediments in palaeo-valleys limit the downward flow of sea water²⁷. Conversely, offshore of North Carolina, palaeo-valleys that have cut through confining layers and are filled with permeable sediments were found to form pathways for seawater intrusion into freshwater aquifers⁶⁷. Therefore, it seems very likely that shelf topography (large width and low gradient) and the presence of relatively thick, low-permeability strata at shallow depths beneath the sea floor played a key part in preserving fresh water by preventing major salinization during the Holocene⁶⁵. However, the fact that low-permeability strata also tend to inhibit the emplacement of fresh water during periods of low sea level provides a conundrum that seems to point at asymmetry in the freshening and salinization parts of the glacial cycles. At glaciated margins such asymmetry may be caused by enhanced lateral freshening owing to ice-sheet influences^{29,48,53}; in some other settings the fact that the major seafloor-confining unit formed during the Holocene and only inhibited the salinization phase³⁵ may explain this. However, in general, the integral glacial cycle response needs to be better understood. Recent high-resolution sampling has revealed a distinctly layered structure of salinity distribution of the New Jersey shelf — fresh water occurring preferentially in thick fine-grained layers³¹ in which diffusion predominates — that may indicate that freshening existed for much longer periods of time than periods of salinization.

Although onshore water tables^{25,26} and offshore salinity gradients⁶⁵ are thought to be the main factors that cause the redistribution of fresh and salt groundwater, other drivers of fluid migration in continental shelf aquifers are known to exist. The presence of high-density brines at relatively shallow depth below the New Jersey shelf³¹ that probably formed by dissolution of salt layers at greater depth, attest to forces that drive fluid flow upward. Potential processes underlying these flows include fluid expulsion due to sediment compaction⁶⁸, and geothermal circulation⁶⁹ due to temperature differences.

Onshore indicators

Despite convincing indications for the widespread presence of offshore palaeo-groundwater, direct observations remain limited (Table 1). However, at many locations onshore hydrogeological and hydrochemical conditions add strong indirect evidence for the presence of fresh groundwater seaward of the coastline^{49,51,70–76}. Radiocarbon dating of groundwater in the onshore part of the VMGRs in Suriname³⁴ and Jakarta⁷⁵ has shown that this water was recharged during the last glacial period. These time constraints are consistent with the inferred conditions that promoted formation of offshore VMGRs (a greater topographic driving force due to lower sea levels and a larger exposed shelf area)².

Fresh coastal groundwater dating back to the last glacial period has further been documented in Florida^{49,74}, Thailand⁵¹, the United Kingdom⁷⁰, Denmark⁷², Portugal⁷¹, Oman⁷⁶ and Tanzania^{17,73} (Fig. 1 and Table 1). Those studies that considered both tracer-based ages and hydrological modelling^{9,50,51,72} confirmed that seaward groundwater flow rates were higher during the glacial period, suggesting that fresh groundwater was driven far beyond the present coastline. Thus, where the offshore geological conditions are conducive to preservation, for example where significant layers of marine clay are found at the sea floor, it can be considered likely that onshore palaeo-groundwater reserves extend under the sea^{70,71}.

Global volume of VMGRs

Two studies^{12,48} have estimated global sub-seafloor freshwater volumes, albeit based on very different methods and for different periods. An estimate of $3 \times 10^5 \text{ km}^3$ of fresh water ($\text{TDS} < 1 \text{ g l}^{-1}$) was reported⁴⁸ based on a volume of 3.8 km^3 fresh water per kilometre of shelf length, as obtained from an interpretation of vertical salinity profiles from the eastern seaboard of the United States and Suriname, and an estimated global continental shelf length of 80,000 km. A much higher estimate of $4.5 \times 10^6 \text{ km}^3$ has been suggested¹² as a possible explanation for elevated ocean salinity during the last glacial maximum that cannot be accounted for solely by the water stored in

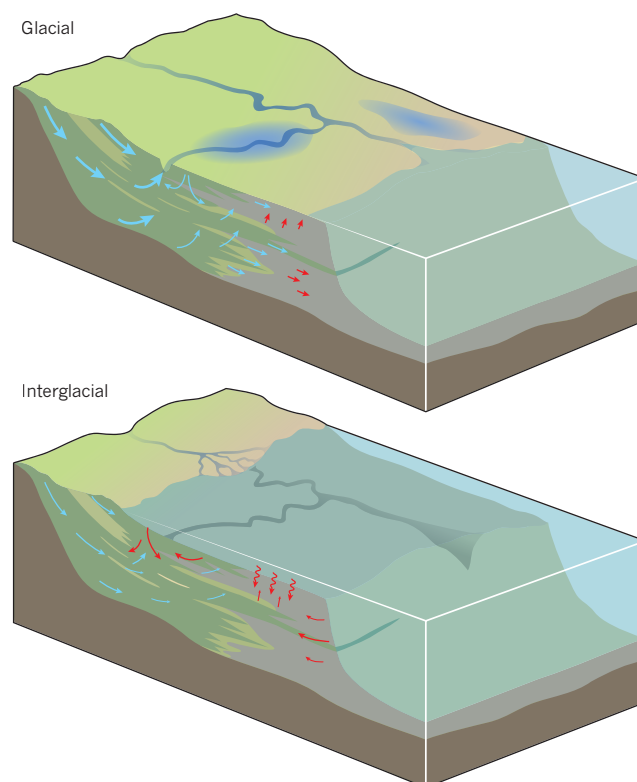


Figure 3 | The geology and the key groundwater flow, and dissolved salt transport processes below the continental shelf. Lower sea levels during glacial periods promote further penetration and recharge of groundwater below continental shelves, whereas incised rivers provide a driving force for topographic flow systems, and saline groundwater retreats seaward. When the shelves are flooded during interglacials, intruded seawater (red arrows) migrates landward as well as downward, while the flow of fresh water (blue arrows) stagnates.

continental ice sheets during that period. To accommodate this large freshwater volume, a continental shelf area corresponding to 5% of the modern ocean surface area would have had to have been flushed to a depth of 500 m with water of negligible salinity when a porosity of 50% is assumed¹². The markedly higher volume estimate for the last glacial maximum compared with the present can partly be attributed to salinization of shelf groundwater by diffusion and density-driven vertical seawater intrusion⁶⁵ owing to sea-level rise. However, modelling has shown that for the massive salinization, implied by the difference between the two volumes, to occur the required timescales approach, or even exceed, the duration of interglacial periods^{8,9,48}. Therefore, the estimate that $4.5 \times 10^6 \text{ km}^3$ of fresh water was sequestered is probably too high, and alternative freshwater stores such as ice sheets need to be considered to be able to explain the observed higher ocean salinity¹² during the last glacial maximum.

Using observational data from Greenland⁵³, Jakarta Bay³⁷ and the Gippsland Basin³⁶ that were not included in the previous study⁴⁸, we estimate the present volume of continental shelf groundwater with a TDS concentration less than 10 g l^{-1} to be $5 \times 10^5 \text{ km}^3$ (Fig. 2 and Box 1). The greatest uncertainty associated with this figure arises from the lack of observational data. There are vast shelf areas, for example the Sunda Shelf in southeast Asia, that were exposed during the last glacial period⁶, but for which there are no groundwater salinity data. New discoveries in these areas could significantly alter the global volume calculations. Despite their very large uncertainty, the calculations show that the volume of fresh and brackish water stored in offshore aquifers may be two orders of magnitude greater than has been extracted globally from continental aquifers since 1900 ($4,500 \text{ km}^3$)¹⁰, and about one-tenth the global volume of shallow

BOX 1

Global volume of VMGRs

The global volume of brackish water ($\text{TDS} < 10 \text{ g l}^{-1}$) stored in offshore VMGRs was estimated by analysing seven shore-normal cross sections (Fig. 2), each of which had a relatively high observation data density so that salinity contour lines could be drawn manually and, for the Nantucket and Greenland cross-sections, derived partly from numerical model simulations. For each cross-section, the volume of brackish water per unit width of shoreline (V_f , $\text{km}^3 \text{ km}^{-1}$) was calculated using:

$$V_f = b \times L \times \phi$$

where b is the brackish water body's average thickness (km) over its shore-normal horizontal width L (km), and ϕ is the aquifer porosity, which was estimated from other studies^{28,33,37,48,61,68,75}. For the complex geometry of the Nantucket cross-section, V_f was calculated as the sum of four individual sub-layers.

Adopted values for b and L , and calculated values of V_f for each cross-section are displayed in Fig. 2. Brackish-water volumes ranged from 1.0 to 9.9 $\text{km}^3 \text{ km}^{-1}$ between the cross-sections, with an average of about 4.5 $\text{km}^3 \text{ km}^{-1}$. Multiplying this number by the total length of passive margins (105,000 km)⁹⁹ yielded a global volume estimate of $5 \times 10^5 \text{ km}^3$. Adopting a threshold TDS concentration of less than 1 g l^{-1} yielded a volume of $3 \times 10^5 \text{ km}^3$. These figures could vary up or down by a factor of about two, owing to the uncertainty of sediment porosity.

(less than 750 m) groundwater ($4.2 \times 10^6 \text{ km}^3$)⁷⁷. These order-of-magnitude calculations suggest that passive margins may represent an important unconventional groundwater resource.

Exploitation

The exploitation of fresh groundwater resources has been pushed beyond sustainable limits in many coastal areas^{19,20,78}, driven by population pressures and increasing economic standards. Moreover, in many coastal cities groundwater extraction from coastal aquifers is inducing considerable and irreversible land subsidence, damaging infrastructure and increasing the incidence of riverine and coastal flooding¹⁸. With more than 40% of the global population within 100 km from the coast⁷⁹, the demand for fresh water will only become more acute in the coming decades, particularly in coastal megacities, and existing problems are likely to be compounded by sea-level rise⁷⁴ and severe drought⁸⁰. Analogous to major inland agricultural areas in India, China and the United States, where current regional strategies are insufficient to address problems with groundwater depletion⁸¹, new or complementary strategies for water provision and management are also required in coastal areas.

As offshore palaeo-groundwaters form on timescales of tens of thousands, if not hundreds of thousands, of years, their production should be considered to be a form of mining, with the same ethical dilemmas as the depletion of non-renewable onshore groundwater resources⁸². Nevertheless, where onshore resources become depleted, exploitation of offshore groundwater could become an option. At the same time, due consideration must be given to the broader spectrum of sustainable water management alternatives, including usage reductions and alternative water sources. Offshore groundwater is not the answer to global water crises, but it has a strategic value that should be acknowledged so that it can be weighed against other options in long-term strategies.

Driven by advances in reverse-osmosis technology, there has been a recent rapid growth of seawater desalination facilities¹⁶ in many coastal regions to augment water supply. The economic feasibility of offshore meteoric groundwater exploitation is therefore best evaluated against

seawater desalination. Desalination costs fluctuate with energy prices, but the costs to desalinate brackish water ($\text{TDS} < 10 \text{ g l}^{-1}$) sourced from VMGRs by reverse osmosis vary between US\$0.10 m^{-3} and \$1.00 m^{-3} , compared with \$0.53 m^{-3} and \$1.50 m^{-3} for sea water¹³. Offshore development necessitates additional infrastructural costs for sea-borne production sites, wells and submarine pipelines¹⁷. These additional investment and operational costs can become prohibitive from an economic perspective; but, if taken into account in the unit water cost, they remain below the higher operational costs for seawater desalination, and the use of offshore low-salinity groundwater may be feasible. If offshore groundwater can be recovered by onshore wells the economics are even more favourable. An example of such a facility already exists in Cape May, New Jersey⁸³, where drinking water has been produced by the desalination of water sourced from an aquifer with an offshore extension since 1998.

Apart from economics, environmental factors need to be considered when investigating the possible use of offshore fresh groundwater¹⁷. Large offshore groundwater abstraction for oil and gas production in Gippsland (Fig. 1), for instance, has seen a considerable drawdown of onshore water tables³⁶. As with seawater desalination, the reject brine from the desalination process needs to be disposed of¹³. Although the use of offshore brackish groundwater has the advantage that the volumes and salinity of the brine are relatively low, disposal will still have an environmental impact. Moreover, the water quality characteristics of groundwater may cause precipitation of salts on the reverse-osmosis membranes, which necessitates the use of chemicals during the production process¹³. From the perspective of drinking water safety, the use of groundwater could mean that concentrations of individual elements such as radium or boron exceed permissible levels¹⁵.

Greater awareness is needed of the adverse impacts of anthropogenic activities on offshore groundwater reserves. The potential of continental shelf aquifers for carbon-dioxide disposal^{36,42} is being assessed. It is quite possible that degradation of offshore VMGRs is already occurring by contamination and enhanced salinization as a result of cross-formational flow along exploration drillholes and wells, or by fluid abstraction for petroleum production³⁶. Moreover, in some areas, offshore groundwater is probably already used, albeit inadvertently, because pumping onshore and on islands draws in groundwater from the offshore parts of aquifers^{8,20,28,32}. The low economic value of water may mean that this is perceived as being of secondary importance at present, but offshore groundwater could prove to be a resource of strategic importance when conventional water management scenarios in coastal areas are no longer adequate or sustainable.

Offshore hydrogeological frontiers

The numerous studies that testify to sub-sea, low-salinity groundwater published^{35–42} since Hathaway and colleagues³² found “anomalous fresh and brackish water” below the New Jersey continental shelf demonstrate that, rather than being an anomaly, low-salinity water below the sea floor is a common phenomenon. It is an expression of the non-stationary nature of the terrestrial hydrological cycle, which spanned a significantly greater surface area throughout much of the Quaternary period compared with the present day, and emphasizes the ever-evolving nature of coastal groundwater systems in response to the dynamics of sea level, landscapes and climate. It also means that addressing the concerns over pressures on coastal water resources, including the adverse effects of predicted sea-level rise^{19,84}, needs to be done with this long-term view in mind; considering future trends as deviations from a static, present-day equilibrium could lead to sub-optimal or even misguided management strategies.

Although the potential benefit of submarine groundwater may form the main impetus for future hydrogeological research of the offshore domain, a better knowledge of groundwater processes under continental shelves will also contribute to the advancement of other fields of research. At present, the links between continental shelf hydrogeology and sub-seafloor ecology and microbiology^{85,86}; material budgets of the oceans, including those of radioisotopes to assess submarine groundwater discharge²²; and seafloor geomorphology² are unclear. The role of

groundwater discharge on exposed continental shelves has even been discussed within the context of the interpretation of the pre-Cambrian fossil record⁸⁷. This places continental shelf hydrogeology at the nexus of other geoscientific disciplines, such as sedimentology, marine geochemistry and reservoir characterization. It is also likely that geochemical and isotopic signatures contained in offshore fresh waters will provide new palaeoclimatic proxies, consistent with such data found in onshore aquifers^{49,88}. Continental shelf hydrogeology could even contribute to advancing our understanding of archaeology, because human settlement and migration patterns may be linked to fresh groundwater discharge zones on exposed continental shelves⁸⁹ in some areas of the world.

Perhaps most importantly from an economic perspective, the circulation of meteoric waters in continental shelf sediments has been found to have an important role in the evolution of sedimentary basins^{90,91} and is key to our understanding of the migration of the oil and gas entrapped in them^{36,41,85}. Improved models of the response of groundwater systems to sea-level variations, as well as the length and timescales associated with meteoric water circulation, can place better constraints on past fluid migration histories.

Although this Review has consolidated evidence for the global occurrence of VMGRs, a paucity of data remains. A wealth of geophysical borehole log data from the hydrocarbon industry probably exists that may still be exploited to better constrain known offshore freshwater occurrences and to reveal numerous unknown ones. Study of these data is both complex and time consuming because of dispersed ownership, data confidentiality and that often most industry borehole measurements only start below the depth at which low-salinity groundwater resources can be expected⁷². New geophysical methods developed for petroleum exploration⁹² also hold great promise for identifying offshore fresh water in continental shelf environments⁹³.

Geophysical methods are of great value but are limited in the sense that they only provide constraints on salinity. More comprehensive data are essential to allow the testing of hypotheses regarding emplacement and preservation. Notably, so far, offshore equivalents of existing onshore studies of noble gases and isotopic tracers^{49,88} have not been carried out, but are much needed to determine the timing and duration of VMGR formation. Improved offshore drilling methods by the International Ocean Discovery Program (formally the Integrated Ocean Drilling Program) have led to better sediment and fluid recovery, allowing detailed profiles of pore-water chemistry⁹⁴ and bringing such studies within reach. Moreover, petrographic and isotopic analysis of diagenetic minerals, which are routinely applied in petroleum reservoir studies to understand fluid migration patterns^{95–97}, have not seen much uptake by the hydrogeological research community, but could be vital to strengthen interpretations about past groundwater flow conditions.

Conversely, hydrogeological studies of present-day groundwater systems could be useful analogues for understanding the relict flow conditions in offshore sedimentary basins, and the preservation potential of offshore freshwater occurrences. This includes onshore areas where vertical seawater intrusion led to aquifer salinization when the coastline was located further inland earlier during the present interglacial period than today⁶⁴. Furthermore, the interpretation of fluid pressures and submarine pore-water chemistries can be aided by numerical modelling of regional-scale groundwater flow^{9,29,48,53}, a technique routinely applied in onshore hydrogeology. From this, it seems clear that scientific advancements can be made when hydrogeologists step across the boundaries of their discipline and team up with other scientists to explore the hidden depths of the continental shelves. ■

Received 20 February; accepted 1 August 2013.

1. Fisher, A. T. Marine hydrogeology: recent accomplishments and future opportunities. *Hydrogeol. J.* **13**, 69–97 (2005).
2. Faure, H., Walter, R. C. & Grant, D. R. The coastal oasis: ice age springs on emerged continental shelves. *Global Planet. Change* **33**, 47–56 (2002).
This article postulates that groundwater discharge and springs were widespread on continental shelves during sea-level low-stands.
3. Lambeck, K. & Chappell, J. Sea level change through the last glacial cycle. *Science* **292**, 679–686 (2001).

4. Clark, P. U. *et al.* The last glacial maximum. *Science* **325**, 710–714 (2009).
5. Soulet, G. *et al.* Glacial hydrologic conditions in the Black Sea reconstructed using geochemical pore water profiles. *Earth Planet. Sci. Lett.* **296**, 57–66 (2010).
6. Voris, H. K. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J. Biogeogr.* **27**, 1153–1167 (2000).
7. Edmunds, W. M. *et al.* in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 289–311 (Geological Society London, 2001).
8. Essaid, H. I. A multilayered sharp interface model of coupled fresh-water and saltwater flow in coastal systems — model development and application. *Wat. Resour. Res.* **26**, 1431–1454 (1990).
9. Harrar, W. G., Williams, A. T., Barker, J. A. & Van Camp, M. in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 213–229 (Geological Society London, 2001).
10. Konikow, L. F. Contribution of global groundwater depletion since 1900 to sea-level rise. *Geophys. Res. Lett.* **38**, L17401 (2011).
11. Lettenmaier, D. P. & Milly, P. C. D. Land waters and sea level. *Nature Geosci.* **2**, 452–454 (2009).
12. Adkins, J. F., McIntyre, K. & Schrag, D. P. The salinity, temperature, and $\Delta^{18}\text{O}$ of the glacial deep ocean. *Science* **298**, 1769–1773 (2002).
13. Greenlee, L. F., Lawler, D. F., Freeman, B. D., Marrot, B. & Moulin, P. Reverse osmosis desalination: water sources, technology, and today's challenges. *Water Res.* **43**, 2317–2348 (2009).
14. Post, V. E. A. Fresh and saline groundwater interaction in coastal aquifers: Is our technology ready for the problems ahead? *Hydrogeol. J.* **13**, 120–123 (2005).
15. Stuyfzand, P. J. & Raat, K. J. Benefits and hurdles of using brackish groundwater as a drinking water source in the Netherlands. *Hydrogeol. J.* **18**, 117–130 (2010).
16. Elimelech, M. & Phillip, W. A. The future of seawater desalination: energy, technology, and the environment. *Science* **333**, 712–717 (2011).
17. Bakken, T. H., Ruden, F. & Mangset, L. E. Submarine groundwater: a new concept for the supply of drinking water. *Water Resour. Manage.* **26**, 1015–1026 (2012).
This is the first article to highlight the potential of submarine groundwater as a source for drinking water.
18. Galloway, D. L. & Burbey, T. J. Regional land subsidence accompanying groundwater extraction. *Hydrogeol. J.* **19**, 1459–1486 (2011).
19. Ferguson, G. & Gleeson, T. Vulnerability of coastal aquifers to groundwater use and climate change. *Nature Clim. Change* **2**, 342–345 (2012).
20. Werner, A. D. *et al.* Seawater intrusion processes, investigation and management: recent advances and future challenges. *Adv. Water Resour.* **51**, 3–26 (2013).
21. Church, T. M. An underground route for the water cycle. *Nature* **380**, 579–580 (1996).
This article discusses the implications of the finding that submarine groundwater discharge is a significant component of the hydrological cycle.
22. Moore, W. S. The effect of submarine groundwater discharge on the ocean. *Annu. Rev. Mar. Sci.* **2**, 59–88 (2010).
23. Taniguchi, M., Burnett, W. C., Cable, J. E. & Turner, J. V. Investigation of submarine groundwater discharge. *Hydrol. Processes* **16**, 2115–2129 (2002).
24. Bratton, J. F. The three scales of submarine groundwater flow and discharge across passive continental margins. *J. Geol.* **118**, 565–575 (2010).
25. Bakker, M. Analytic solutions for interface flow in combined confined and semi-confined, coastal aquifers. *Adv. Water Resour.* **29**, 417–425 (2006).
26. Kooi, H. & Groen, J. Offshore continuation of coastal groundwater systems: predictions using sharp-interface approximations and variable-density flow modelling. *J. Hydrol.* **246**, 19–35 (2001).
This was the first study to provide quantitative constraints on the offshore extension of active submarine groundwater discharge.
27. Krantz, D. E., Manheim, F. T., Bratton, J. F. & Phelan, D. J. Hydrogeologic setting and ground water flow beneath a section of Indian River Bay, Delaware. *Ground Water* **42**, 1035–1051 (2004).
28. Johnston, R. H. The salt-water–fresh-water interface in the tertiary limestone aquifer, southeast Atlantic outer continental-shelf of the USA. *J. Hydrol.* **61**, 239–249 (1983).
29. Person, M. *et al.* Pleistocene hydrogeology of the Atlantic continental shelf, New England. *Geol. Soc. Am. Bull.* **115**, 1324–1343 (2003).
30. Malone, M. J., Claypool, G., Martin, J. B. & Dickens, G. R. Variable methane fluxes in shallow marine systems over geologic time — the composition and origin of pore waters and authigenic carbonates on the New Jersey shelf. *Marine Geology* **189**, 175–196 (2002).
31. van Geldern, R. *et al.* Stable isotope geochemistry of pore waters and marine sediments from the New Jersey shelf: methane formation and fluid origin. *Geosphere* **9**, 96–112 (2013).
This study demonstrates previously unrecognized salinity stratification based on high-resolution pore-water data from the New Jersey continental shelf.
32. Hathaway, J. C. *et al.* United-States geological survey core drilling on the Atlantic shelf. *Science* **206**, 515–527 (1979).
This is the seminal paper that demonstrated the widespread occurrence of low-salinity groundwater below the continental shelf of the eastern United States.
33. Person, M. *et al.* Use of a vertical $\Delta^{18}\text{O}$ profile to constrain hydraulic properties and recharge rates across a glacio-lacustrine unit, Nantucket Island, Massachusetts, USA. *Hydrogeol. J.* **20**, 325–336 (2012).
34. Groen, J., Post, V. E. A., Kooi, H. & Hemker, C. J. in *Tracers and Modelling in*

- Hydrogeology* (ed. Dassargues, A.) 417–424 (2000).
35. Groen, J., Velstra, J. & Meesters, A. Salinization processes in paleowaters in coastal sediments of Suriname: evidence from $\Delta^{7}\text{Cl}$ analysis and diffusion modelling. *J. Hydrol.* **234**, 1–20 (2000).
 36. Varma, S. & Michael, K. Impact of multi-purpose aquifer utilisation on a variable-density groundwater flow system in the Gippsland Basin, Australia. *Hydrogeol. J.* **20**, 119–134 (2012).
 37. Maathuis, H., Mak, W. & Adi, S. in *Groundwater: Past Achievements and Future Challenges* (ed. Sililo, O.) 209–213 (Balkema, 2000).
 38. Oteri, A. U. Electric log interpretation for the evaluation of salt water intrusion in the eastern Niger Delta. *Hydro. Sci. J.* **33**, 19–30 (1988).
 39. Grasby, S. E., Chen, Z., Issler, D. & Stasiuk, L. Evidence for deep anaerobic biodegradation associated with rapid sedimentation and burial in the Beaufort-Mackenzie basin, Canada. *Appl. Geochem.* **24**, 536–542 (2009).
 40. Zhang, Z., Zou, L., Cui, R. & Wang, L. Study of the storage conditions of submarine freshwater resources and the submarine freshwater resources at north of Zhoushan sea area. *Marine Sci. Bull.* **30**, 47–52 (2011).
 41. Davies, C. P. N. *Hydrocarbon evolution of the Bredasdorp basin, Offshore South Africa: from Source to Reservoir*. PhD thesis, Univ. Stellenbosch (1997).
 42. Hennig, A. & Otto, C. A Hydrodynamic Characterisation of the Offshore Vlaming Sub-basin. (CO₂CRC, 2005).
 43. Post, V. E. A., Hooijboer, A. E. J., Groen, J., Gieske, J. M. J. & Kooi, H. in *Proc. 16th Salt Water Intrusion Meeting, Wolin Island, Poland* (ed. Sadurski, A.) (SWIM, 2000).
 44. Kriete, C., Suckow, A. & Harazim, B. Pleistocene meteoric pore water in dated marine sediment cores off Callao, Peru. *Estuar. Coast. Shelf Sci.* **59**, 499–510 (2004).
 45. Expedition 317 Scientists. Site U1353. *Proc. Integr. Ocean Drill. Program* **317**, 103 (2011).
 46. Middelburg, J. J. & de Lange, G. J. The isolation of Kau Bay during the last glaciation: direct evidence from interstitial water chlorinity. *Neth. J. Sea Res.* **24**, 615–622 (1989).
 47. Meisler, H., Leahy, P. P. & Knobel, L. L. *Effect of Eustatic Sea-Level Changes on Saltwater–Freshwater relations in the Northern Atlantic coastal plain*. (U.S. Geological Survey, 1984).
 48. Cohen, D. et al. Origin and extent of fresh paleowaters on the Atlantic Continental Shelf, USA. *Ground Water* **48**, 143–158 (2010).
 49. Morrissey, S. K., Clark, J. F., Bennett, M., Richardson, E. & Stute, M. Groundwater reorganization in the Floridan aquifer following Holocene sea-level rise. *Nature Geosci.* **3**, 683–687 (2010).
 50. Love, A. J. et al. Groundwater residence time and paleohydrology in the Otway basin, south Australia — H-2, O-18 and C-14 data. *J. Hydrol.* **153**, 157–187 (1994).
 51. Sanford, W. E. & Buapeng, S. Assessment of a groundwater flow model of the Bangkok basin, Thailand, using carbon-14-based ages and paleohydrology. *Hydrogeol. J.* **4**, 26–40 (1996).
 52. Robb, J. M. Spring sapping on the lower continental slope, offshore New Jersey. *Geology* **12**, 278–282 (1984).
 53. DeFoor, W. et al. Ice sheet-derived submarine groundwater discharge on Greenland's continental shelf. *Water Resour. Res.* <http://dx.doi.org/10.1029/2011WR010536> (28 July 2011).
 54. Mann, R. & Gieskes, J. M. Interstitial water studies, Leg 28. Initial Rep. *Deep Sea Drill. Proj.* **28**, 805–814 (1975).
 55. Chambers, S. R. Solute distributions and stable isotope chemistry of interstitial waters from Prydz Bay, Antarctica. *Proc. Ocean Drill. Program* **119**, 375–392 (1991).
 56. Edmunds, W. M. in *Isotopes in the Water Cycle: Past, Present and Future of a Developing Science*, 341–352 (Springer, 2005).
 57. De Carlo, E. H. Geochemistry of pore water and sediments recovered from the Exmouth Plateau. *Proc. Ocean Drill. Program* **122**, 295–308 (1992).
 58. Kastner, M., Elderfield, H. & Martin, J. B. Fluids in convergent margins — what do we know about their composition, origin, role in diagenesis and importance for oceanic chemical fluxes? *Phil. Trans. R. Soc. A* **335**, 243–259 (1991).
 59. Kastner, M. et al. Diagenesis and interstitial-water chemistry at the Peruvian continental margin; major constituents and strontium isotopes. *Proc. Ocean Drill. Program* **112**, 413–440 (1990).
 60. Mora, G. Isotope-tracking of pore water freshening in the fore-arc basin of the Japan Trench. *Mar. Geol.* **219**, 71–79 (2005).
 61. Gieskes, J. M., Lawrence, J. R. & Galleis, G. Interstitial water studies, Leg 38. Initial Rep. *Deep Sea Drill. Proj.* **38–41**, 121–133 (1978).
 62. Exon, N. F. et al. Leg 189 Summary. *Proc. Ocean Drill. Program* **189**, 1–98 (2001).
 63. Hesse, R. Pore water anomalies of submarine gas-hydrate zones as tool to assess hydrate abundance and distribution in the subsurface — What have we learned in the past decade? *Earth-Science Reviews* **61**, 149–179 (2003).
 64. Post, V. E. A. & Kooi, H. Rates of salinization by free convection in high-permeability sediments: insights from numerical modeling and application to the Dutch coastal area. *Hydrogeol. J.* **11**, 549–559 (2003).
 65. Kooi, H., Groen, J. & Leijnse, A. Modes of seawater intrusion during transgressions. *Wat. Resour. Res.* **36**, 3581–3589 (2000).
 - This was the first study to evaluate the modes of salinization of continental shelf aquifers during sea-level rise.**
 66. Hughes, J. D., Vacher, H. L. & Sanford, W. Temporal response of hydraulic head, temperature, and chloride concentrations to sea-level changes, Floridan aquifer system, USA. *Hydrogeol. J.* **17**, 793–815 (2009).
 67. Mulligan, A. E., Evans, R. L. & Lizarralde, D. The role of paleochannels in groundwater/seawater exchange. *J. Hydrol.* **335**, 313–329 (2007).
 68. Dugan, B. & Flemings, P. B. Overpressure and fluid flow in the New Jersey continental slope: Implications for slope failure and cold seeps. *Science* **289**, 288–291 (2000).
 69. Wilson, A. M. The occurrence and chemical implications of geothermal convection of seawater in continental shelves. *Geophys. Res. Lett.* **30**, 2127 (2003).
 70. Edmunds, W. M. et al. in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 71–92 (Geological Society London, 2001).
 71. Condeso de Melo, M. T., Carreira Paqueta, P. M. & Marques da Silva, M. A. in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 139–154 (Geological Society London, 2001).
 72. Hinsby, K. et al. in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 29–48 (Geological Society London, 2001).
 73. Bakari, S. S. et al. Groundwater residence time and paleorecharge conditions in the deep confined aquifers of the coastal watershed, South-East Tanzania. *J. Hydrol.* **466–467**, 127–140 (2012).
 74. Sanford, W. E. Groundwater hydrology coastal flow. *Nature Geosci.* **3**, 671–672 (2010).
 75. Geyh, M. A. & Sofner, B. Groundwater analysis of environmental carbon and other isotopes from the Jakarta basin aquifer, Indonesia. *Radiocarbon* **31**, 919–925 (1989).
 76. Weyhenmeyer, C. E. et al. Cool glacial temperatures and changes in moisture source recorded in Oman groundwaters. *Science* **287**, 842–845 (2000).
 77. Berner, E. K. & Berner, R. A. *Global Water Cycle: Geochemistry and Environment*. 397 (Prentice Hall, 1987).
 78. Post, V. & Abarca, E. Saltwater and freshwater interactions in coastal aquifers. *Hydrogeol. J.* **18**, 1–4 (2010).
 79. Martínez, M. L. et al. The coasts of our world: ecological, economic and social importance. *Ecol. Econ.* **63**, 254–272 (2007).
 80. Appleyard, S. J., Angeloni, J. & Watkins, R. Arsenic-rich groundwater in an urban area experiencing drought and increasing population density, Perth, Australia. *Appl. Geochem.* **21**, 83–97 (2006).
 81. Aeschbach-Hertig, W. & Gleeson, T. Regional strategies for the accelerating global problem of groundwater depletion. *Nature Geosci.* **5**, 853–861 (2012).
 82. van der Gun, J. & Lipponen, A. Reconciling groundwater storage depletion due to pumping with sustainability. *Sustainability* **2**, 3418–3435 (2010).
 83. Barlow, P. M. *Ground Water in Freshwater–Saltwater Environments of the Atlantic Coast* (US Geological Survey, 2003).
 84. Green, T. R. et al. Beneath the surface of global change: impacts of climate change on groundwater. *J. Hydrol.* **405**, 532–560 (2011).
 85. Berndt, C. Focused fluid flow in passive continental margins. *Phil. Trans. R. Soc. A* **363**, 2855–2871 (2005).
 86. Schippers, A. et al. Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature* **433**, 861–864 (2005).
 87. Xiao, S. & Knauth, L. P. Fossils come in to land. *Nature* **493**, 28–29 (2013).
 88. Loosli, H. H. et al. in *Palaeowaters in Coastal Europe: Evolution of Groundwater Since the Late Pleistocene*, Vol. 189 (eds Edmunds, W. M. & Milne, C. J.) 193–212 (Geological Society London, 2001).
 89. Bailey, G. N. & King, G. C. P. Dynamic landscapes and human dispersal patterns: tectonics, coastlines, and the reconstruction of human habitats. *Quat. Sci. Rev.* **30**, 1533–1553 (2011).
 90. Morad, S., Ketzer, J. M. & De Ros, L. F. Spatial and temporal distribution of diagenetic alterations in siliciclastic rocks: implications for mass transfer in sedimentary basins. *Sedimentology* **47**, 95–120 (2000).
 91. Screaton, E. J. Recent advances in subseafloor hydrogeology: focus on basement-sediment interactions, subduction zones, and continental slopes. *Hydrogeol. J.* **18**, 1547–1570 (2010).
 92. Constable, S. & Srnka, L. J. An introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration. *Geophysics* **72**, WA3–WA12 (2007).
 93. Hoefel, F. G. & Evans, R. L. Impact of low salinity porewater on seafloor electromagnetic data: a means of detecting submarine groundwater discharge? *Estuar. Coast. Shelf Sci.* **52**, 179–189 (2001).
 94. Mountain, G. S., Proust, J. N., McInroy, D. & the Expedition 313 scientists in *Proc. IODP 313* (IODP, 2009).
 95. Mansurbeg, H. et al. Meteoric-water diagenesis in late Cretaceous canyon-fill turbidite reservoirs from the Espírito Santo Basin, eastern Brazil. *Mar. Pet. Geol.* **37**, 7–26 (2012).
 96. Lundegard, P. D. & Trevena, A. S. Sandstone diagenesis in the Pattani basin (Gulf of Thailand) — history of water rock interaction and comparison with the Gulf of Mexico. *Appl. Geochem.* **5**, 669–685 (1990).
 97. Bazin, B., Brosse, E. & Sommer, F. Chemistry of oil-field brines in relation to diagenesis of reservoirs 1: use of mineral stability fields to reconstruct in situ water composition. Example of the Mahakam basin. *Mar. Pet. Geol.* **14**, 481–495 (1997).
 98. Amante, C. & Eakins, B. W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis* 19 (NOAA, 2009).
 99. Bradley, D. C. Passive margins through earth history. *Earth Sci. Rev.* **91**, 1–26 (2008).

Author Information Reprints and permissions information is available at www.nature.com/reprint. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/pmeydy. Correspondence should be addressed to V.P. (vincent.post@flinders.edu.au).

Ecosystem-based coastal defence in the face of global change

Stijn Temmerman¹, Patrick Meire¹, Tjeerd J. Bouma², Peter M. J. Herman², Tom Ysebaert^{2,3} & Huib J. De Vriend⁴

The risk of flood disasters is increasing for many coastal societies owing to global and regional changes in climate conditions, sea-level rise, land subsidence and sediment supply. At the same time, in many locations, conventional coastal engineering solutions such as sea walls are increasingly challenged by these changes and their maintenance may become unsustainable. We argue that flood protection by ecosystem creation and restoration can provide a more sustainable, cost-effective and ecologically sound alternative to conventional coastal engineering and that, in suitable locations, it should be implemented globally and on a large scale.

Coastal flood disasters are an ever-present threat to coastal societies. Recent examples include the flooding caused by Hurricane Katrina in 2005 in New Orleans, Cyclone Nargis in 2008 in southern Myanmar, Hurricane Sandy in 2012 in New York, and Typhoon Haiyan last month in the central Philippines. Such flood disasters are caused by extreme storm surges that can raise the local sea level by several metres through severe wind, waves and atmospheric pressure conditions¹. Coastal flood risks are likely to increase over the coming decades owing to global and regional changes that include increasing storm intensity^{2,3}, accelerating sea-level rise and land subsidence⁴ (Fig. 1). Growing coastal populations mean more people will be exposed to these increasing flood risks⁵. At least 40 million people and US\$3,000 billion of assets are located in flood-prone coastal cities today, and these are expected to increase to 150 million people and \$35,000 billion by 2070 (ref. 5) (Fig. 2).

Conventional coastal engineering, such as the building of sea walls, dykes and embankments, is widely perceived as the ultimate solution to combat flood risks. However, these defences are seriously challenged in many locations as their continual and costly maintenance, as well as their heightening and widening to keep up with the increasing flood risk are becoming unsustainable. Furthermore, conventional coastal engineering often exacerbates land subsidence by soil drainage⁴ and hinders the natural accumulation of sediments by tides, waves and wind, thereby compromising the natural adaptive capacity of shorelines to keep up with relative sea-level rise (Fig. 1).

In recent years, ecosystem-based flood defence has been brought into large-scale practice as a regional solution that is more sustainable and cost-effective than conventional coastal engineering. It is applied at locations that have sufficient space between urbanized areas and the coastline to accommodate the creation of ecosystems, such as tidal marshes, mangroves, dunes, coral reefs and shellfish reefs, that have the natural capacity to reduce storm waves^{6–8} and storm surges^{9–11}, and can keep up with sea-level rise by natural accretion of mineral and biogenic sediments^{12,13} (Fig. 1). The latter process secures the long-term sustainability of ecosystem-based coastal protection. Furthermore, these ecosystems provide several added benefits¹⁴, including water quality improvement, fisheries production and recreation, so that in the long term they could be more cost effective than conventional defences^{15,16} (Table 1). This ecosystem-based approach is not suitable for all coastal areas and its

global application is still scarce. On the basis of current knowledge, drawn largely from tidal wetland creation projects, we argue that the approach has the potential to protect many of the world's largest flood-prone coastal populations (Fig. 2).

Challenges to conventional coastal engineering

During past centuries, wetlands in river deltas and estuaries were reclaimed on a large scale and turned into rich agricultural, urban and industrial areas such as New York, New Orleans, Shanghai, Tokyo and, on a country scale, the Netherlands. Consequently, today's deltas and estuaries are host to the world's largest flood-prone coastal populations⁵ (Fig. 2) and have lost most of their natural flood defences.

Wetland reclamation leads to the loss of storage area for flood waters so that storm surges rise higher and propagate faster and further inland through the remaining channels of a delta or estuary (Fig. 1). For example, in the inland part of the Scheldt estuary, Belgium, high water levels have increased by 1.3 m since 1930, which is about five times faster than the rise of high water levels at the coast¹⁷. This landward amplification of rising high water levels is exacerbated by extensive wetland reclamation (which diminishes the flood storage area and reduces resistance to landward flood propagation) and by channel dredging (which further facilitates flood propagation)¹¹. Similar effects have been observed in other engineered estuaries and may rapidly increase in Asia, for instance, where deltaic wetlands are being reclaimed and channels engineered on large scales⁴.

In addition, as reclaimed wetlands are cut off from the sea or estuary, the natural process of sediment deposition and land rise is inhibited (Fig. 1). Decreased wetland sedimentation may also result from reductions in river-borne sediment supply by upstream river dams, river diversions and the building of embankments between the river and wetlands. This has contributed to large-scale wetland submergence as the sea level rises, for example, in the Venice lagoon¹⁸ and the Mississippi delta¹⁹. The increasing difference between sea and land levels is further exacerbated by soil subsidence due to compaction, soil drainage and extraction of groundwater, oil and gas⁴. For instance, subsidence over the twentieth century amounts to 5 m in Tokyo, 3 m in Shanghai, and 2 m in Bangkok⁵. In the Netherlands subsidence has resulted in 9 million people living below mean sea level²⁰.

¹Ecosystem management research group, University of Antwerp, Antwerp 2610, Belgium. ²Spatial ecology research group, Royal Netherlands Institute for Sea Research, Yerseke 4400 AC, the Netherlands. ³Institute for Marine Resources and Ecosystem Studies (IMARES), Yerseke 4400 AB, the Netherlands. ⁴Delft University and EcoShape, Dordrecht 3311 JG, the Netherlands.

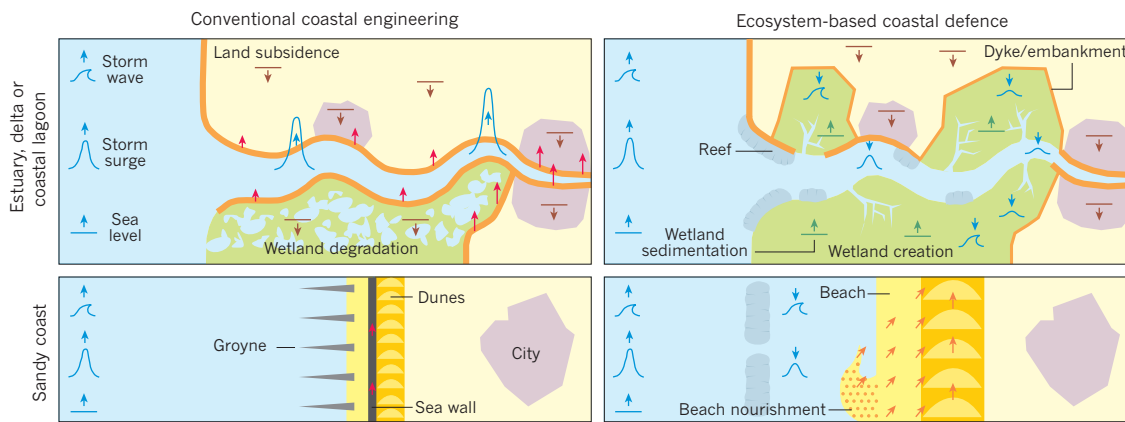


Figure 1 | Conventional coastal engineering compared with new ecosystem-based defence. The schematic maps illustrate global and regional changes that increase the risk of coastal flood disasters (blue arrows indicate an increase or decrease in intensity of storm waves, storm surge and sea level), and the basic principles of flood protection by conventional coastal engineering (left) and new ecosystem-based defences (right) for an estuary, delta or coastal lagoon (top) and a sandy coast (bottom). In the case of conventional defences, red arrows indicate the need for maintenance and heightening of dykes, embankments and sea walls with sea-level rise. In an engineered estuary, delta or coastal lagoon (top left), embankment of wetlands stimulates the landward

heightening of storm surges and exacerbates land subsidence (brown arrows) due to inhibited sediment supply and soil drainage. In the case of ecosystem-based defence in an estuary, delta or coastal lagoon (top right), wetland and reef creation attenuate landward storm surge propagation and storm waves, and stimulate wetland sedimentation (green arrows) with sea-level rise. For an engineered sandy coast (bottom left), groynes and sea walls may provoke dune degradation due to hindered sand supply, whereas for ecosystem-based defence along a sandy coast (bottom right), reefs help to attenuate storm waves and surge, and offshore sand nourishment stimulates beach and dune sedimentation with sea-level rise (orange arrows).

Conventional engineering solutions — hard structures such as sea walls and embankments — are used to protect today's populated and vulnerable coastlines, but they are seriously challenged by ever rising maintenance costs and unwanted ecological side effects. For example, adjustments to the Dutch flood defence system that are required to cope with increasing flood risks, are expected to cost up to €1.6 billion per year by 2050, whereas the potential damage from insufficient defence may amount to €3,700 billion²⁰. Conventional engineering of deltas and estuaries has culminated in projects such as the Delta Works in southwest Netherlands or the Thames barrier in London, where complete estuaries were cut off from the sea by permanent dams or movable storm surge barriers. However, as the Dutch example has shown, these projects can have serious economic and ecological drawbacks, including the erosion of tidal habitats and the occurrence of toxic algal blooms that kill aquatic life in the enclosed or semi-enclosed estuaries²¹. In the case of sandy coasts, conventional coastal engineering — such as groynes and sea walls — can block the wave- and wind-driven supply of sand, thereby compromising the long-term build-up of beaches and dunes with rising sea levels (Fig. 1).

New ecosystem-based flood defence

The creation or restoration of large coastal ecosystems provides a new alternative or add-on to conventional coastal defences, as coastal ecosystems attenuate storm waves^{6–8} and surges^{9–11}, and accumulate sediments with sea-level rise^{12,13}. The viability of different ecosystem-based approaches depends on the type of coastal area and location of the city at risk (Fig. 1).

For cities located in estuaries or deltas — such as New Orleans, London and many large Asian cities (dark green and pale green in Fig. 2) — the creation or restoration of large tidal marshes or mangroves between the city and the sea provides extra water storage areas and friction, which attenuates the landward propagation of storm surges^{9–11} and reduces flood risks in the densely populated hinterland. Local marsh restoration for natural habitat conservation is reasonably widespread in several countries, but large-scale, estuarine-wide implementation of wetland restoration mainly for flood defence is very limited. In the Belgian Scheldt estuary up to 4,000 ha of historically reclaimed wetlands are being converted back into floodplains, of which about 2,500 ha will become tidal marshes²². The first marsh was created in 2006 and the total

project should be completed by 2030 at an expected cost of around €600 million¹⁶. By comparison, the yearly risk of flood damage is estimated at €1 billion¹⁶ by 2100, if this flood defence project is not implemented. The marshes are created by landward displacement of historical dykes (Fig. 1) or by building sluices through the dykes that allow tidal flooding and marsh development on the previously reclaimed land^{23,24} (Fig. 3). Similar projects are being deployed in UK estuaries, such as in the Humber estuary, where conversion of historically reclaimed land into marshes for coastal defence is called 'managed coastal realignment'¹⁵. In the United States several large projects are underway, including the restoration of around 8,000 ha of tidal marshes in San Francisco Bay, California²⁵. And, in the Mississippi delta, large marshlands have been restored to protect New Orleans from hurricane flooding on the basis of studies reporting that every kilometre of marshland reduces hurricane surge levels by 5 to 10 cm^{9,19}. Degraded marshes are restored with dredged sediment and by diverting the sediment-laden Mississippi water back into the delta¹⁹. In tropical regions, such as southeast Asia, mangrove forest plantations are being considered as protection against storm surges²⁶. A recent study in Florida has shown that hurricane surge level can be reduced by 40 to 50 cm per kilometre of mangrove forest width¹⁰.

For cities behind sandy coastlines (yellow in Fig. 2), such as Amsterdam, Abidjan in the Ivory Coast and Lagos in Nigeria, beach and dune barriers are crucial defences against coastal flooding. As part of the Building with Nature programme in the Netherlands — a combined initiative between national authorities, dredging contractors, engineering consultants and research institutes — a large hook-shaped sand peninsula has been created by depositing 21 million cubic metres of sand on the shoreface off the coast of Holland²⁷. The aim of this project is to combat coastal erosion along a 17-km stretch of coastline over several decades. This is achieved by the natural distribution of artificially deposited sand by tide, wave and wind force towards beaches and dunes. This approach avoids the need for more conventional beach nourishment, whereby local habitat is disturbed by the frequent direct deposition of sand onto beaches. The creation of oyster reefs is another example of an ecosystem-based coastal defence project. These reefs have been constructed from gabions filled with oyster shells on eroding sand flats in the Eastern Scheldt estuary²⁸. Oyster larvae attach themselves to the

shells in the summer to form robust, living oyster reefs that reduce waves, currents and erosion²⁹ and at the same time generate other services, such as essential fish habitat.

Potentials and limitations

Ecosystem-based flood defence has several additional benefits compared with conventional engineering approaches, including the improvement of water quality, carbon sequestration, the production of fisheries, nature conservation and the creation of recreational space (Table 1). For example, tidal wetlands improve the water quality of estuaries by delivering scarce nutrients such as silica³⁰ and by acting as a sink for abundant nutrients such as nitrogen and contaminants such as heavy metals³¹. This improvement in water quality suppresses the growth of toxic algae and stimulates phytoplankton growth, which is essential for the food web. Mangroves and marshes are important sinks for atmospheric CO₂ and therefore contribute to climate change mitigation³². However, the experience in the United Kingdom indicates that it may take several decades before created tidal marshes function in a similar biogeochemical way to naturally occurring marshes³³. Wetlands and reefs promote fisheries production by providing an indispensable habitat for juvenile fish, shellfish and crustaceans¹⁴. Natural sedimentation ensures that ecosystem-based projects are better self-sustained when exposed to sea-level rise in the long term²⁴. Tidal wetlands may become submerged, however, in regions where the tidal range and the sediment supply are critically low and subsidence rates are high¹². Ecosystem-based projects can be more cost effective in the long run than continued conventional defence. A cost–benefit analysis for the Humber estuary, UK, revealed that after 25 years tidal marsh restoration on reclaimed land is economically more beneficial than maintaining dykes¹⁵. A similar study for the Scheldt estuary that compared conventional engineering with marsh creation concluded that an ecosystem-based approach is more cost effective and will have recovered the initial costs of implementation after a period of 20 years¹⁶.

However, there are important limitations to the wider implementation of ecosystem-based flood defences (Table 1). These defences tend to require more space than conventional structures. In particular, in highly urbanized coastal areas, such as New York or Tokyo, space is so scarce that only conventional coastal engineering seems to be feasible. The more space that is available between the sea and the urbanized areas at risk, the higher the efficiency of ecosystem-based flood defence. This applies to several of the world's largest flood-prone cities located far inland in deltas or estuaries (Fig. 2). In these locations, large mangroves or marshes between the sea and the city could significantly contribute to storm surge protection (Fig. 1). For cities closer to the mouth of estuaries and deltas, a combination of ecosystem-based and engineered defences would be more appropriate. Where little space is available on land, seaward ecosystem creation — such as off-shore reef creation — could be an option. However, projects should ensure that they do not destroy or disturb the existing sequence of ecosystems along the sea-to-land transition, including off-shore reefs, intertidal flats and wetlands.

At present, the nature of the created ecosystem and its effectiveness for flood defence remains, in part, uncertain because too few long-term studies exist. So far, most of our experience is of tidal marsh creation. In locations with low site elevation and high tidal inundation frequency and duration, the growth of marsh vegetation is not necessarily successful³⁴. This could also be a problem for the creation of mangrove forests; however, even less data exists for tropical regions^{35,36}. Therefore, wetland-creation sites should be carefully selected on the basis of suitable elevations. If only low elevation sites are available, tidal inundation can be reduced with the help of hydraulic structures such as weirs or sluices (Fig. 3) — as has been applied in the Scheldt estuary^{23,24} — or the sites can be raised with dredged sediments or river-borne sediment supply, such as in the Mississippi delta¹⁹. This, of course, increases the overall cost, but the cost–benefit balance is still positive — as for the Scheldt estuary project¹⁶, for which sluices are built to create tidal marshes

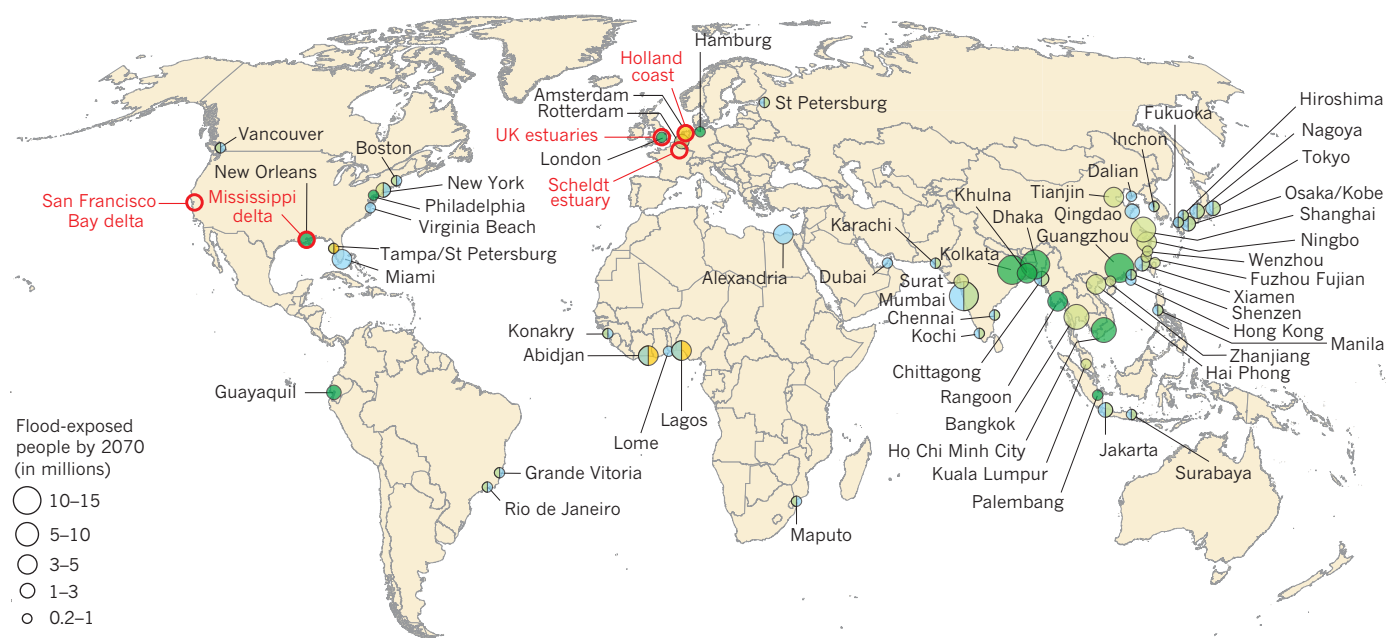


Figure 2 | Global need for coastal flood protection, and large-scale examples and potential application of ecosystem-based defence. The map shows only cities with more than 200,000 people exposed to coastal flood risks by 2070 worldwide (predictions are based on data from ref. 5). We classified all cities into four categories according to the potential application of ecosystem-based defence: cities in estuaries or deltas and more than 50 km from the sea (dark green) can be well-protected from flooding by marshes or mangroves and moderately protected by reefs, in addition to

conventional engineering; cities in estuaries or deltas but less than 50 km from the sea (pale green) can be moderately protected by marshes or mangroves and by engineering, and somewhat protected by reefs; cities more than 5 km from the sea and behind a sandy coast (orange) can be well protected by dunes and protected to some extent by engineering; cities right at the coast (blue) can be protected by engineering and to some extent by reefs. Existing examples of large-scale applications of ecosystem-based flood defence are shown in red.

Table 1 | Potential and limitations of conventional compared with ecosystem-based coastal defence

Affected variable	Conventional coastal engineering	Ecosystem-based coastal defence
Natural habitat	Degradation or destruction	Conservation or restoration
Sediment accumulation (after sea-level rise)	Disturbed or stopped by embankments, groynes, dams, and so on.	Sustained (if enough sediment is available)
Land subsidence	Exacerbated by wetland reclamation, soil drainage, groundwater and gas extraction	Counterbalanced by sediment trapping, but continues behind ecosystems
Storm surge propagation through an estuary or delta	Wetland reclamation reduces water storage and friction, enhancing inland storm surges	Wetland restoration enlarges water storage and friction, lowering inland storm surges
Long-term sustainability	Low: regular maintenance is needed at high cost	High: ecosystems are self-maintaining (if enough sediment is available)
Cost–benefit appraisal	Moderate to high	Mostly high due to added benefits
Water quality of estuary, delta and coastal sea	May degrade by organic matter accumulation and toxic algal growth in closed-off estuaries	Improved and sustained by nutrient and contaminant cycling in restored wetlands
Climate mitigation through carbon sequestration	None	Mangroves and marshes are important carbon sinks
Fisheries and aquaculture production	Reduced: less habitat for young fish, shellfish and crustaceans due to wetland reclamation	Improved: more habitat for young fish, shellfish and crustaceans due to wetland and reef restoration
Human recreation potential	Negative perception of artificial landscape	Positive perception of natural landscape
Required space	Moderate	High, therefore, not applicable for cities on the coast
Difficulty of creating the defence structure	Moderate	Relatively high due to natural dynamics and variability
Existing implementation and experience	Substantial, but many failures in the past	Limited so far. More research is urgently needed
Social and political acceptance	Widely accepted	So far, only accepted in certain areas (Europe and United States)
Health hazards (other than flooding)	None	Wetlands with stagnant water may facilitate breeding of mosquitoes that could spread disease

on low elevation sites (Fig. 3). Soil properties and their effect on water logging, wind waves and biotic factors such as seed dispersal, bioturbation and grazing may also hamper wetland development^{34,35}. To maximize the success of ecosystem creation, a step-wise implementation is advised: starting with small-scale pilot projects with intensive interdisciplinary monitoring and expanding these to large-scale projects with a suitably adjusted design and implementation²³. Absolute success of ecosystem-based flood defences cannot be guaranteed, but this is also true for conventional flood defences.

Public perception may seriously hinder the realization of large-scale ecosystem-based flood defence. The idea that valuable land, laboriously reclaimed by previous generations, should be turned back into wetlands could provoke considerable public opposition. For example, although Belgium and the United Kingdom are converting reclaimed land into tidal marshes on a large scale, there is much more social and political objection to similar projects in the Netherlands, where more than 50% of the Dutch population live below sea level and struggle

against the water is strongly embedded in the nation's cultural heritage. The examples from Belgium and the United Kingdom show that societal opposition can be overcome by clear communication of the benefits of ecosystem-based defences between scientific, political and governmental institutions, local stakeholders and the general public. Other obstacles for public acceptance may include the fear that wetlands facilitate mosquito breeding and disease transmission, especially in subtropical or tropical areas. Socio-economic factors must also be considered and may partly explain why ecosystem-based defence on a large scale has only been implemented in Europe and the United States and on a much smaller scale in Asia. Despite the considerable need for flood protection and the high potential for ecosystem-based solutions in many areas of Asia (Fig. 2).

Wider implementation

Global and regional changes have forced us to search for sustainable adaptation strategies to protect against coastal flood hazards. Ecosystem-based strategies can be used to remedy the limitations of continued conventional engineering in suitable coastal settings, particularly in deltas that host the world's largest flood-prone populations. Recent implementations of these strategies demonstrate that ecosystem-based flood defence can be more sustainable and cost-effective than conventional flood defences, with additional benefits and with fewer side effects. These findings should further stimulate joint research by ecologists and engineers, and motivate governments and industry to support the wider implementation of ecosystem-based flood defence. ■

Received 3 March; accepted 18 September 2013.

- Resio, D. T. & Westerink, J. J. Modelling the physics of storm surges. *Phys. Today* **61**, 33–38 (2008).
- Knutson, T. R. *et al.* Tropical cyclones and climate change. *Nat. Geosci.* **3**, 157–163 (2010).
- Lin, N., Emanuel, K., Oppenheimer, M. & Vanmarcke, E. Physically based assessment of hurricane surge threat under climate change. *Nature Clim. Change* **2**, 462–467 (2012).
- Syvitski, J. P. M. *et al.* Sinking deltas due to human activities. *Nat. Geosci.* **2**, 681–686 (2009).

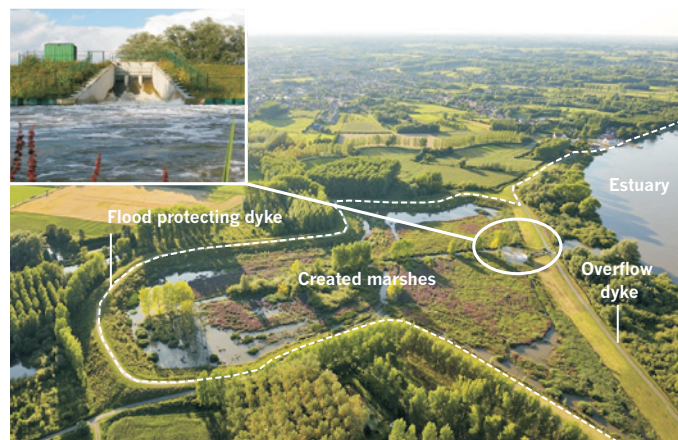


Figure 3 | Ecosystem-based flood defence. A man-made marsh in the Scheldt estuary, Belgium, protects more landward, densely populated areas from storm surge flooding. The sluice (inset) allows daily tidal flooding of the marsh.

5. Nicholls, R. J. *et al.* *Ranking of the World's Cities Most Exposed to Coastal Flooding Today and in the Future* (OECD, 2007).
This report provides a comprehensive global assessment of people and assets in coastal cities exposed to storm surge flood risks in 2005 and expected by 2070.
6. Barbier, E. B. *et al.* Coastal ecosystem-based management with nonlinear ecological functions and values. *Science* **319**, 321–323 (2008).
This paper highlights the wave attenuation function of several coastal ecosystem types, and that this protective function is nonlinearly related to ecosystem size.
7. Gedan, K. B., Kirwan, M. L., Wolanski, E., Barbier, E. B. & Silliman, B. R. The present and future role of coastal wetland vegetation in protecting shorelines: answering recent challenges to the paradigm. *Clim. Change* **106**, 7–29 (2011).
8. Shepard, C. C., Crain, C. M. & Beck, M. W. The protective role of coastal marshes: a systematic review and meta-analysis. *PLoS ONE* **6**, e27374 (2011).
9. Wamsley, T. V., Cialone, M. A., Smith, J. M., Atkinson, J. H. & Rosati, J. D. The potential of wetlands in reducing storm surge. *Ocean Eng.* **37**, 59–68 (2010).
10. Zhang, K. Q. *et al.* The role of mangroves in attenuating storm surges. *Estuar. Coast. Shelf Sci.* **102–103**, 11–23 (2012).
11. Temmerman, S., De Vries, M. B. & Bouma, T. J. Coastal marsh die-off and reduced attenuation of coastal floods: a model analysis. *Global Planet. Change* **92–93**, 267–274 (2012).
12. Kirwan, M. L. *et al.* Limits on the adaptability of coastal marshes to rising sea level. *Geophys. Res. Lett.* **37**, L23401 (2010).
13. Fagherazzi, S. *et al.* Numerical models of salt marsh evolution: ecological, geomorphic, and climatic factors. *Rev. Geophys.* **50**, RG1002 (2012).
14. Barbier, E. B. *et al.* The value of estuarine and coastal ecosystem services. *Ecol. Monogr.* **81**, 169–193 (2011).
15. Turner, R. K., Burgess, D., Hadley, D., Coombes, E. & Jackson, N. A cost-benefit appraisal of coastal managed realignment policy. *Glob. Environ. Change* **17**, 397–407 (2007).
16. Broekx, S., Smets, S., Liekens, I., Bulckaen, D. & De Nocker, L. Designing a long-term flood risk management plan for the Scheldt estuary using a risk-based approach. *Nat. Hazards* **57**, 245–266 (2011).
17. Temmerman, S., Govers, G., Wartel, S. & Meire, P. Modelling estuarine variations in tidal marsh sedimentation: response to changing sea level and suspended sediment concentrations. *Mar. Geol.* **212**, 1–19 (2004).
18. Carniello, L., Defina, A. & D'Alpaos, L. Morphological evolution of the Venice lagoon: evidence from the past and trend for the future. *J. Geophys. Res.* **114**, F04002 (2009).
19. Day, J. W. *et al.* Restoration of the Mississippi delta: lessons from hurricanes Katrina and Rita. *Science* **315**, 1679–1684 (2007).
20. Kabat, P. *et al.* Dutch coasts in transition. *Nat. Geosci.* **2**, 450–452 (2009).
21. Verspagen, J. M. H. *et al.* Water management strategies against toxic *Microcystis* blooms in the Dutch delta. *Ecol. Appl.* **16**, 313–327 (2006).
22. Sigma Plan. <http://www.sigmaplan.be/en> (Sigmaplan, 2011).
23. Maris, T. *et al.* Tuning the tide: creating ecological conditions for tidal marsh development in a flood control area. *Hydrobiologia* **588**, 31–43 (2007).
24. Vandenbruwaene, W. *et al.* Sedimentation and response to sea-level rise of a restored marsh with reduced tidal exchange: comparison with a natural tidal marsh. *Geomorphology* **130**, 115–126 (2011).
25. United States Environmental Protection Agency. *San Francisco Bay Delta Watershed* <http://www2.epa.gov/sfbay-delta> (EPA, 2013).
26. Schmitt, K., Albers, T., Pham, T. T. & Dinh, S. C. Site-specific and integrated adaptation to climate change in the coastal mangrove zone of Soc Trang Province, Vietnam. *J. Coast. Conserv.* **17**, 545–558 (2013).
27. van Slobbe, E. *et al.* Building with nature: in search of resilient storm surge protection strategies. *Nat. Hazards* **65**, 947–966 (2013).
28. Ecoshape. *Oyster reefs for tidal flat protection in the Eastern Scheldt* <http://www.ecoshape.nl/oyster-reefs-eastern-scheldt.html> (Ecoshape, 2012).
29. Scyphers, S. B., Powers, S. P., Heck, K. L. & Byron, D. Oyster reefs as natural breakwaters mitigate shoreline loss and facilitate fisheries. *PLoS ONE* **6**, e22396 (2011).
30. Struyf, E., Temmerman, S. & Meire, P. Dynamics of biogenic Si in freshwater tidal marshes: Si regeneration and retention in marsh sediments (Scheldt estuary). *Biogeochemistry* **82**, 41–53 (2007).
31. Teuchies, J., Beauchard, O., Jacobs, S. & Meire, P. Evolution of sediment metal concentrations in a tidal marsh restoration project. *Sci. Total Environ.* **419**, 187–195 (2012).
32. Mcleod, E. G. L. *et al.* A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO₂. *Front. Ecol. Environ.* **9**, 552–560 (2011).
33. Burden, A., Garbutt, R. A., Evans, C. D., Jones, D. L. & Cooper, D. M. Carbon sequestration and biogeochemical cycling in a saltmarsh subject to coastal managed realignment. *Estuar. Coast. Shelf Sci.* **120**, 12–20 (2013).
34. Wolters, M., Garbutt, A. & Bakker, J. P. Salt-marsh restoration: evaluating the success of de-embankments in north-west Europe. *Biol. Conserv.* **123**, 249–268 (2005).
35. Friess, D. A. *et al.* Are all intertidal wetlands naturally created equal? Bottlenecks, thresholds and knowledge gaps to mangrove and saltmarsh ecosystems. *Biol. Rev. Camb. Philos. Soc.* **87**, 346–366 (2012).
36. Balke, T. *et al.* Seedling establishment in a dynamic sedimentary environment: a conceptual framework using mangroves. *J. Appl. Ecol.* **50**, 740–747 (2013).

Acknowledgments We gratefully acknowledge financial support to our research that is related to this article by the Research Foundation — Flanders (FWO), the Research Fund of the University of Antwerp (BOF), Waterwegen & Zeekanaal NV, the EU-funded THESUES project (FP7.2009-1, Contract 244104), STW grant 07324, the Singapore–Delft Water Alliance, the innovation program Building with Nature and its contributors (including the Netherlands Ministry of Infrastructure and the Environment, the European Fund for Regional Development, the Municipality of Dordrecht and the partners of the EcoShape consortium).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/riej27. Correspondence should be addressed to S.T. (stijn.temmerman@uantwerpen.be).

Green and golden seaweed tides on the rise

Victor Smetacek¹ & Adriana Zingone²

Sudden beaching of huge seaweed masses smother the coastline and form rotting piles on the shore. The number of reports of these events in previously unaffected areas has increased worldwide in recent years. These ‘seaweed tides’ can harm tourism-based economies, smother aquaculture operations or disrupt traditional artisanal fisheries. Coastal eutrophication is the obvious, ultimate explanation for the increase in seaweed biomass, but the proximate processes that are responsible for individual beaching events are complex and require dedicated study to develop effective mitigation strategies. Harvesting the macroalgae, a valuable raw material, before they beach could well be developed into an effective solution.

Green, brown and red seaweeds lying on the beach are part and parcel of life in many coastal regions. The amount of beached seaweed biomass started to increase along the shores of industrialized countries in the 1970s, and by the 1990s had become a nuisance along many beaches^{1,2}, when mass-stranding events of macroalgae became known as green tides. During the 2000s the number of reports from new locations all over the world increased further, as did the magnitude of the beaching events³. Although non-toxic to humans, seaweed tides harm shore-based activities by virtue of their sheer physical mass. Tonnes of seaweed smothering the shoreline deter tourists and the dense, drifting seaweeds can prevent swimmers and small boats from accessing the sea (Fig. 1); if not removed in time, the algae can turn into a stinking morass, which can produce toxic hydrogen sulphide (H₂S) from its anoxic interior^{1,3}, and have major detrimental effects on the affected coastal ecosystems^{1,2,4–6}.

Surprisingly, the extensive seaweed tides are mainly the result of only a few genera of macroalgae. Two genera are especially prominent. Species of the genus *Ulva*, which now includes the former genus *Enteromorpha*⁷, are mainly responsible for green tides. The thallus (vegetative body) is only one or two cells thick but the shapes vary even within species and can be sheet-like, tubular or fern-shaped⁸. *Sargassum* — from which the Sargasso Sea takes its name — is the other genus; and we suggest the term ‘golden tide’ to describe the massive shoaling events it is responsible for (after the apt description of floating *Sargassum* as “The golden floating rainforest of the Atlantic Ocean”⁹). The *Sargassum* thallus is leathery, tough and differentiated into features that resemble leaves and a stem, and has well-developed gas bladders for flotation. Both *Ulva* and *Sargassum* are cosmopolitan, exceptionally species-rich genera and increase their growth rate in response to nutrients^{10,11}. Whereas most species will only grow when attached to a hard substrate, a few can substantially increase their biomass in a free-floating state, either by increasing the size of the thalli and their fragments, or by making new floating thalli. This is crucial (discussed later) because it is the unattached forms that, by invading new space (the water column), are able to increase their nutrient supply, free themselves from competition for limited hard substrates and avoid their many benthic grazers. As a result, unattached forms can build up large biomasses, forming the massive seaweed tides we discuss in this Perspective. Green tides have occurred all over the world, whereas golden tides have been restricted to beaches between the Gulf of Mexico and Bermuda; however, they significantly increased their range during a spectacular 2011 event.

Ulva green tides

The increase in *Ulva* biomass on European and US beaches that began in the 1970s was linked to coastal eutrophication¹. As the many harmful effects became evident, the countries affected took measures to reduce nutrient input to the sea from agricultural sources and sewage. A decline in nutrient concentrations resulted in abatement of the problem in the southern North Sea¹². In other regions, particularly along the popular tourist beaches of Brittany, the magnitude of green tides has been increasing since the 1970s¹³. Beached seaweed has traditionally been collected and used as fertilizer by local farmers, but by the 1990s it had to be taken away by the truckload (Fig. 1). In 2009, H₂S gas from *Ulva* rotting on a Brittany beach caused the death of a horse, and, in 2011, the death of around 30 wild boars. Both incidents were widely reported in the press with some headlines giving the impression that the algae were toxic. The resulting effect on tourism caused severe losses to the local economy, in addition to the costs of removing and disposing the 100,000 tonnes of beached algae (estimated to be US\$10–150 per tonne)¹³.

Increases in the accumulation of *Ulva* biomass coincided with the expansion of factory livestock farming in Brittany. The consensus among the scientific community seems to be that eutrophication from the effluents of intensive stock rearing is the cause of the increase in number and magnitude of green tides since the 1990s¹³. The meat-producing and tourist industries are both mainstays of the provincial economy, and, following the animal deaths, confrontation between the two increased¹⁴. Brittany is a wet region overloaded with nutrients released by the high density of animals — equivalent to those from 50 million people¹³ — and so eutrophication is inevitable because the manure is not being shipped back to the animal feed producers outside the province. In efforts to make the best out of a situation that is unlikely to change soon, *Ulva* biomass has been used as a raw material for biogas production, organic fertiliser and as an additive to animal and human food. However, the value barely meets the costs of current methods of algal collection and processing¹³.

In 2008, a spectacular green tide invaded, without warning, the beaches of Qingdao — the venue for the sailing events of the Beijing Olympics. Masses of *Ulva* floated in from the open water of the Yellow Sea and beached a few weeks before the competition was due to start, ensuring prominent coverage by the international media (Fig. 1). A 30-km-long boom was deployed to keep the masses of floating algae out of the bay, and the removal of more than a million tonnes of algae

¹Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany. ²Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy.

from the beaches involved 10,000 people at an estimated cost to the province of \$30 million¹⁵. In addition, aquaculture operations along the shore suffered losses of \$100 million³.

Given the novelty of the phenomenon, the origin, genesis and relationship to eutrophication of the green tide was traced with exemplary speed¹⁶. Analysis of satellite images of the Yellow Sea in 2008 revealed that in total 3,500 km² were covered with floating algae in patches spread over 84,000 km². The high nitrate concentrations in the Yellow Sea could explain the high growth rates (21.9% per day) of the floating algal patches^{15–17}. These patches were subsequently driven inshore by winds, hitting a 140 km stretch of coastline, which included Qingdao. The species responsible, *Ulva prolifera*, was subsequently shown to proliferate rapidly by sporulation of cells of fragmented thalli; these grew into new thalli by attaching to the parent thallus, with 1–2-mm diameter fragments having the highest sporulation rate¹⁸.

The pelagic seaweed bloom, as well as those in subsequent years, could be traced in satellite images to the coastline some 200 km south of Qingdao where aquaculture of the edible red alga *Porphyra yezoensis*, which is grown on rafts along the intertidal zone, has expanded rapidly since 2004 (refs 17, 19). Because *Ulva prolifera* grows profusely on the rafts, thallus fragments dislodged and discarded in the sea during harvesting of *Porphyra* in spring are the most likely seed source of the mid-summer green tide. If this seeding hypothesis is correct, then collecting and using *Ulva* as a by-product would mitigate the magnitude of the green tide. It is estimated that 500 tonnes of *Ulva* thalli, discarded from the *Porphyra* rafts, grow into one million tonnes in 6 weeks¹⁵. Another hypothesis based on genetic signatures (5S rDNA spacer sequences), suggests that the *Ulva* strain responsible for the green tide could be overwintering on the sediment surface south of the Yellow Sea as fragments that stem from the summer surface bloom²⁰. The hypotheses are

not necessarily mutually exclusive, as the swarm of seeding fragments moving northward from the *Porphyra* rafts could be augmented by thalli fragments rising to the surface from shallow sediments (Fig. 2). Whatever the source of seeding thalli, green tides along the coasts of the Yellow Sea are recurrent, with the 2013 event reportedly reaching a record level²¹.

Managing green tides

The case for a direct connection between spreading coastal eutrophication and the worldwide upsurge in the incidence of green tides is compelling^{1,2}. However, curbing eutrophication requires significant investment in infrastructure and agricultural practices in the catchment area and can take years to implement, and even longer to take effect. Thus, although water quality in Tokyo Bay has improved, *Ulva* green tides have increased; the species responsible overwinters as unattached thalli drifting at the sediment surface of shallow waters²². The notorious *Ulva* blooms of Venice in the late 1980s, however, are no longer such a nuisance^{23,24}, even though nutrient concentrations — in particular nitrate — have not significantly diminished in the subsequent decades²⁴. As biomass accumulation is a function of the seed population multiplied by its growth rate, spreading seed banks of overwintering, free-floating strains of local *Ulva* could obviate the effects of reducing eutrophication because they are protected from the many *Ulva* grazers (such as snails and crustaceans) that live on the sea floor^{25,26}. Thus, green tides have been a chronic problem in the eutrophied northern Baltic since the 1970s — unattached *Ulva intestinalis* thalli overwinter in shallow, ice-covered waters and rise in the water column, commencing growth in spring²⁷. Free-floating *Ulva* species generally persist from spring until autumn in a growth state, after which they transfer the sequestered nutrients to the sediments. In

CRISTINA BARROCA



MAX FRELING



ANSA/JIAN FENG



ANDREW HUCKBODY



Figure 1 | Green and golden tides. A moderate *Ulva* green tide on a beach in Brittany, France, (top left) and the Qingdao beach, China, during a green tide (top right). Clean up of *Sargassum* golden tide in a bay in Antigua, southern Caribbean (bottom left) and a golden tide on a beach in Sierra Leone, Africa during the spectacular 2011 event (bottom right).

shallow, enclosed seas and fjords these are returned to the surface by vertical mixing (that is, retained within the system)¹³. Thus, in favourable topographical and hydrological environments, free-floating macroalgae are likely to continue to proliferate and maintain the eutrophic state that is favourable for their perpetuation.

The occurrence and magnitude of green tides often vary both annually and seasonally, hence the planning of mitigation measures will require interdisciplinary investigation of the life-cycle strategies of the species involved in relation to the physical and chemical setting of the environment: topography, circulation, wind patterns and the nutrient regime, exemplified by the work done on the Yellow Sea green tides^{3,15–20,28–30}. An obvious mitigation target would be the overwintering and early spring growing stages. The costs of collecting and disposing of *Ulva* masses in spring would have to be weighed against their nuisance 'value' in the summer. In the Yellow Sea, the profit from *Porphyra* aquaculture amounts to \$53 million, whereas the cost of removing *Ulva* from the beaches is estimated to be \$30 million¹⁵. As elsewhere, the overburdened local governing bodies are responsible for keeping their beaches clean. In Brittany, the tourism industry (\$5.1 billion) and the farming sector (\$11.6 billion) have been pitted against each other^{31,32}. Recently, policy makers at national and European Union levels have passed measures to curb factory farming practices, which has led to closures, lay-offs and protests in the area³³. Similar tensions have begun

to arise in China between the Jiangsu province, home to *Porphyra* aquaculture, and Shandong province, whose coasts have been affected by green tides³⁴.

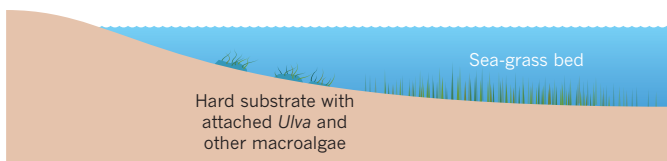
Sargassum golden tides

Golden tides due to the beaching of floating *Sargassum* occur regularly in summer along the coasts of the Gulf of Mexico and are often a nuisance on tourist beaches³⁵. An increase in golden tides during the 1980s and 1990s has been linked to higher nutrient loads of the Mississippi river¹⁰. However, compared with *Sargassum* in the Sargasso Sea — often referred to as "the only sea without a coastline"⁹ — little is known about *Sargassum* in the Gulf of Mexico. Analysis of satellite images from 2002 to 2008 revealed that floating *Sargassum* originated in the north-western Gulf of Mexico each spring and was exported to the Sargasso Sea where it accumulated in the summer months and by winter had disappeared³⁶, presumably because aged thalli sank to the deep sea³⁷. From satellite images, an estimated one million tonnes wet weight of *Sargassum* is exported to the Atlantic each year³⁶. For comparison, about the same mass of *Ulva* was collected and disposed of on land during the green tide of the 2008 Olympics³, and a similar amount was estimated to have accumulated in the Venice lagoon during the peak outbreak²³.

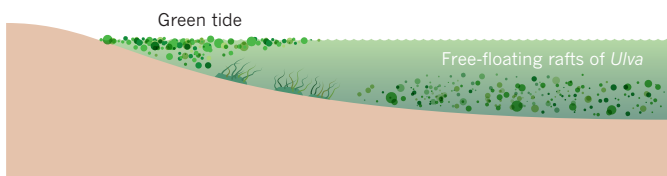
Floating *Sargassum*, represented by the two species *Sargassum natans* and *Sargassum fluitans*, is deemed a valuable and unique habitat that harbours many highly adapted, and even endemic animal species that depend on *Sargassum* for food⁹. If *Sargassum* is imported to the Sargasso Sea each year, it is hard to imagine how the dependent animal species, which beach with the *Sargassum*, could have evolved with such a 'one-way' seasonal life cycle. Nevertheless, the satellite observations are supported by ship-based reports of the seasonal cycle of floating *Sargassum*³⁶. Thus, the processes maintaining the floating *Sargassum* habitat through the winter and the specific properties of the north-western Gulf that allow rapid growth of new thalli during spring, warrant investigation. *S. natans* and *S. fluitans* are considered to be the only holopelagic macroalgae, so why is a specific geographical seeding site part of their life cycle? This question gains importance because of the commercial interest in *Sargassum* biomass; patents have been filed for growing and harvesting *Sargassum* in the Sargasso Sea. Furthermore, an international alliance of scientists has recently been formed to protect and manage the Sargasso Sea⁹.

During 2011, there was an ocean-scale build-up of *Sargassum* that at its peak extended across the Atlantic and resulted in massive golden tides along the west African coast, from Sierra Leone to Ghana, and, on the other side of the Atlantic, from Trinidad to the Dominican Republic (Fig. 3)³⁸. The peak biomass during the 2011 event was 200-fold higher than the previous 8 years' average biomass peak recorded in the region³⁹. According to eyewitnesses, beached *Sargassum* was unknown in northwest Africa before 2011, so the event came as a shock to the many afflicted fishing villages, and has been attributed to the effects of offshore oil production that had started at the time⁴⁰. The afflicted southern Caribbean islands had never, within living memory, experienced an event of this magnitude³⁸. Satellite images showed that the algal rafts had developed along the northern coast of Brazil, north of the mouth of the Amazon, from where they moved east and west, eventually stretching from shore to shore (Fig. 3)³⁶. The effects on the beaches were substantial; however, because no popular tourist beaches or big cities were affected the African events were only reported in regional media. Along the western coast of Ghana a blanket of *Sargassum* extended for kilometres offshore, clogging fishing nets and impeding the passage of small boats. This resulted in food shortages for people living in villages dependent on artisanal fisheries for their livelihood⁴¹. In the Caribbean, tourism was affected because of the closure of beaches and bays (Fig. 1). Unfortunately, the satellite sensor from which the images in Fig. 3 were produced (MERIS) went out of service in early 2012 (ref. 39). We could not find any reports of unusually large golden tides in subsequent years from either the southern Caribbean islands or the west coast of Africa.

a Pristine state



b Eutrophied state



c Green tide

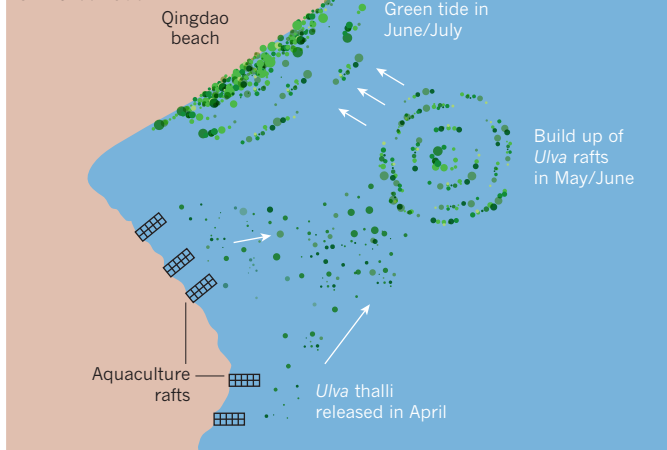


Figure 2 | *Ulva* green tide development in a shallow coastal environment. a, The pristine state with sea-grass beds. b, The eutrophic state with thick, free-floating mats of *Ulva* drifting above the sea bed; these smother sea-grass beds and are subsequently deposited as green tides on adjoining beaches. c, The Yellow Sea green tide originated from *Ulva prolifera* growing on aquaculture rafts 200 km south of Qingdao in April. The algae built up on the surface of the Yellow Sea during May and June and were deposited on the beaches in June and July.

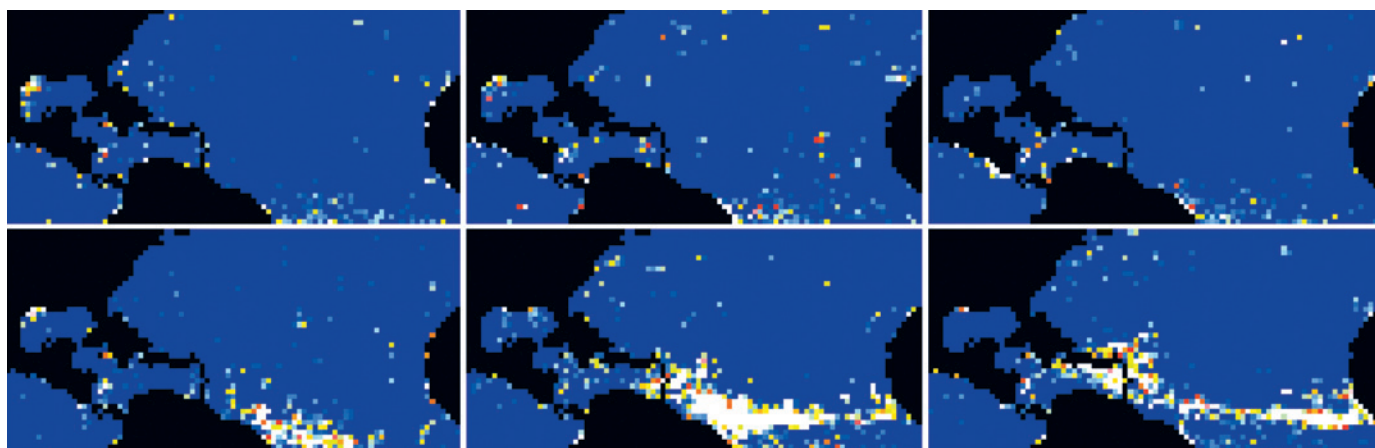


Figure 3 | Distribution of drifting *Sargassum* rafts derived from MERIS satellite images across the central Atlantic Ocean. An average year (2010, top panels) compared with the spectacular 2011 event (bottom panels). From left to right, panels show May, July and September for

both years. High concentrations are shown in white and red, and low concentrations in deep blue. Note the high *Sargassum* concentrations in the north-western Gulf of Mexico in May for both years. Reprinted with permission from ref. 39.

Explaining the unprecedented, ocean-scale build-up of *Sargassum* biomass in 2011 will require much detective work by physical, chemical and biological oceanographers because its occurrence challenges the current concept that the Sargasso Sea is a closed system⁹. Indeed, an unanswered question raised in the 1970s asked why there are five subtropical ocean gyres (STG) but only one, the North Atlantic, harbours the Sargasso Sea. The circulation pattern of all STGs is essentially similar and all impinge on extensive coastlines along their western flanks. The long, unbroken evolutionary history of floating *Sargassum* is evidenced by the many adaptations, in particular perfect camouflage, that various animal classes have evolved in response to life in this floating habitat. The highly specialized sargassum fish (*Histrio histrio*) has a pan-tropical distribution (including the west African coast)⁴², prompting the questions: does *Sargassum* exist in large enough quantities to provide a habitat for specialized species in other STGs and how could this change in the future. Was the 2011 *Sargassum* bloom a freak event caused by a unique collection of non-linear environmental factors impinging on each other to create an environment in which *Sargassum* thrived? Or does it represent a symptom of an ocean-wide response to increasing pressure on the biosphere from anthropogenic waste, and hence an indication of what is to come?

Establishing an international consortium

Blooms of noxious phytoplankton (unicellular microalgae), known as red tides and later as harmful algal blooms (HABs), occur in coastal regions worldwide. In most cases, the harmful effects are caused by phytoplankton species that render seafood toxic, result in the mass death of marine animals or affect aquaculture operations. High-biomass phytoplankton blooms can also be a nuisance on beaches by discolouring the water (red and brown tides) or forming scums and foams⁴³. Their biomass is low compared with green tides, but they can also cause anoxic events, albeit in deeper water. HAB research has substantially increased over the past decades, greatly profiting from international organisations such as the Global Ecology and Oceanography of Harmful Algal Blooms programme (IOC–SCOR GEOHAB)⁴⁴. The seaweed blooms discussed in this Perspective are very different² and only have two features in common with microalgal HABs: they grow in a free-floating state (that is, they compete with phytoplankton for nutrients) and cause harm to the affected coastlines. Because they are essentially benthic organisms that live a planktonic existence, the investigation of free-floating macroalgae will need to combine established research methods from both fields, and apply new techniques for surveying, sampling and modelling them. It would be advisable to develop such a dedicated, interdisciplinary research programme at an international level.

The focus of this scientific network could be the comparatively few

species of macroalgae that can all increase biomass in a free-floating stage, either drifting above the bottom in shallow waters, or floating at the surface further out at sea. Most green-tide species belong to the former category^{1,2} and, to our knowledge, the latter has only been reported in the Yellow Sea. The surface-floating *Ulva prolifera* strain has been shown to differ from attached *Ulva* species of the Yellow Sea coastline¹⁵. Is this surface-floating form a new genotype that can only attain massive biomass in settings with the hydrographic features and wind patterns of the Yellow Sea? Or are other shallow seas, such as the North Sea, susceptible to invasion by a surface-floating form of *Ulva*? This could happen either by evolution within the local species pool or by inadvertent introduction of the Yellow Sea form. Among the questions that need to be addressed is the buoyancy regulation mechanism of free-floating *Ulva*, which allows thalli to stay suspended in the water column or rise to the surface. Furthermore, life cycles also need to be studied *in situ*. The ability to produce new thalli on the surface of the parent thallus¹⁸ suggests the absence of chemical deterrents, which are known to exist for the thallus surface of other *Ulva* species^{45,46}. In the case of *Sargassum*, the factors necessary for large-scale spring regeneration of thalli require elucidation, as well as hindcasting for the factors that allowed the 2011 *Sargassum* event. Finally, the fact that surface-floating *Ulva* and *Sargassum* rafts were able to proliferate in the open sea indicates their ability to compete with phytoplankton for nutrients. Biomass build-up of dense macroalgal clumps, in contrast to diffuse phytoplankton, is presumably enabled by wind energy pushing the rafts through the water, which can vastly increase their nutrient supply. One wonders why evolution of this surface-floating macroalgal life form in the open sea is restricted to so few genera.

Mitigation or amelioration?

An in-depth understanding of the growth dynamics of massive seaweed tide species is not only a prerequisite for developing cost-effective mitigation strategies, but it could also provide the basic knowledge required to manage free-floating algae as a potentially valuable resource. *Ulva* biomass contains a number of compounds of interest to the food-additive industry, and a biorefinery plant to process the *Ulva* biomass collected from the beach or shallow water has recently been established in the region of Brittany plagued by green tides⁴⁷. One of the aims is to provide an alternative food additive to the fish-meal-based one currently given to farmed fish. In order for this to work, the costs of collecting *Ulva* need to be similar to those of collecting fish used to make fish meal. In the case of surface-floating *Ulva*, ‘catching’ the algae at sea by making use of ships from the current fishing fleet, which are already equipped with the facilities required to preserve the catch, should be competitive in terms of cost. Patches of floating rafts located by aerial surveys could be concentrated with booms similar to those used for containing oil spills, and the thalli

pumped on deck and collected on nets and filters of decreasing mesh size (to collect fragments). If carried out in the early stages of the bloom, this technique could mitigate the magnitude of the beaching events. In other regions, techniques to collect *Ulva* masses in shallow water from special ships by rakes, nets or suction pipes could be developed and one technique is already in operation in the Venice lagoon²³. Transporting the *Ulva* biomass to processing factories from harbours should be much cheaper and easier than using rakes and tractors on beaches; the quality of the raw material will also be superior because it will be fresher and contain less sand. Price depends on demand, so encouraging the establishment of *Ulva*-processing factories will raise the price of the raw material — and could well make floating rafts of *Ulva* an interesting target for a new, summer fishery. In the case of *Sargassum*, also a valuable raw material, harvesting by ship is already regulated in the western Atlantic⁴⁸.

Needless to say, harvesting floating macroalgae is the logical and ultimate step in the process known as 'fishing down marine food webs'⁴⁹. It should also be pointed out that the carbon-to-nitrogen ratio of oceanic *Sargassum* is around 50:1 (ref. 10) and the alga's tendency to rapidly sink to the deep-sea floor³⁷ makes it a much more efficient vehicle to artificially sequester carbon in the oceans than phytoplankton, which have carbon-to-nitrogen ratios of less than 10:1. The future impact of green and golden tides could be very different if they become regarded as potential crops rather than harmful weeds. ■

Received 13 September, accepted 25 October 2013.

- Fletcher, R. T. in *Marine Benthic Vegetation - Recent Changes and the Effects of Eutrophication* (eds Schramm, W. & Nienhuis, P. H.) 7–43 (Springer, 1996).
- Valiela, I. et al. Macroalgal blooms in shallow estuaries: controls and ecophysiological and ecosystem consequences. *Limnol. Oceanogr.* **42**, 1105–1118 (1997).
- Ye, N. H. et al. 'Green tides' are overwhelming the coastline of our blue planet: taking the world's largest example. *Ecol. Res.* **26**, 477–485 (2011).
- Norkko, A. & Bonsdorff, E. Population responses of coastal zoobenthos to stress induced by drifting algal mats. *Mar. Ecol. Prog. Ser.* **140**, 141–151 (1996).
- Norkko, A. & Bonsdorff, E. Rapid zoobenthic community responses to accumulations of drifting algae. *Mar. Ecol. Prog. Ser.* **131**, 143–157 (1996).
- Arroyo, N. L., Aarnio, K., Mäensivu, M. & Bonsdorff, E. Drifting filamentous algal mats disturb sediment fauna: Impacts on macro-mesofaunal interactions. *J. Exp. Mar. Biol. Ecol.* **420–421**, 77–90 (2012).
- Hayden, H. S. et al. Linnaeus was right all along: *Ulva* and *Enteromorpha* are not distinct genera. *Eur. J. Phycol.* **38**, 277–294 (2003).
- Blomster, J. et al. Novel morphology in *Enteromorpha* (*Ulvoephyceae*) forming green tides. *Am. J. Bot.* **89**, 1756–1763 (2002).
- Laffoley, D. A. et al. *The Protection and Management of the Sargasso Sea: The Golden Floating Rainforest of the Atlantic Ocean* 1–44 (Washington, 2011).
- Lapointe, B. E. A comparison of nutrient-limited productivity in *Sargassum natans* from neritic vs. oceanic waters of the western North Atlantic Ocean. *Limnol. Oceanogr.* **40**, 625–633 (1995).
- Teichberg, M. et al. Eutrophication and macroalgal blooms in temperate and tropical coastal waters: nutrient enrichment experiments with *Ulva* spp. *Glob. Change Biol.* **16**, 2624–2637 (2010).
- van Beusekom, J. E. E. et al. *Quality Status Report 2009. Wadden Sea Ecosystem No. 25* (eds Marencic, H. & de Vlas, J.) 1–21 (Common Wadden Sea Secretariat, Trilateral Monitoring and Assessment Group, 2009).
- Charlier, R. H., Morand, P. & Finkl, C. W. How Brittany and Florida coasts cope with green tides. *Int. J. Environ. Stud.* **65**, 191–208 (2008).
- Saltmarsh, M. A battle between economic mainstays in Brittany. <http://www.nytimes.com/2010/08/21/business/energy-environment/21toxic.html?pagewanted=all> (New York Times, 2010).
- Liu, D. et al. The world's largest macroalgal bloom in the Yellow Sea, China: formation and implications. *Estuar. Coast. Shelf Sci.* **129**, 2–10 (2013).
- Sun, S. et al. Emerging challenges: Massive green algae blooms in the Yellow Sea. *Nature Preced.* <http://dx.doi.org/10.101/npre.2008.2266.1> (2008).
- Keesing, J. K., Liu, D., Fearn, P. & Garcia, R. Inter- and intra-annual patterns of *Ulva prolifera* green tides in the Yellow Sea during 2007–2009, their origin and relationship to the expansion of coastal seaweed aquaculture in China. *Mar. Pollut. Bull.* **62**, 1169–1182 (2011).
- Gao, S. et al. A strategy for the proliferation of *Ulva prolifera*, main causative species of green tides, with formation of sporangia by fragmentation. *PLoS ONE* **5**, e8571 (2010).
- Hu, C. et al. On the recurrent *Ulva prolifera* blooms in the Yellow Sea and East China Sea. *J. Geophys. Res.* **115**, C05017 (2010).
- Liu, F. et al. Understanding the recurrent large-scale green tide in the Yellow Sea: temporal and spatial correlations between multiple geographical, aquacultural and biological factors. *Mar. Environ. Res.* **83**, 38–47 (2013).
- Jacobs, A. With surf like turf, huge algae bloom befouls China coast. <http://www.nytimes.com/2013/07/06/world/asia/>
- huge-algae-bloom-afflicts-qingdao-china.html (New York Times, 2013).
- Yabe, T. et al. Green tide formed by free-floating *Ulva* spp. at Yatsu tidal flat, Japan. *Limnology* **10**, 239–245 (2009).
- Sfriso, A. & Marcomini, A. Decline of *Ulva* growth in the lagoon of Venice. *Bioresour. Technol.* **58**, 299–307 (1996).
- Facca, C., Pellegrino, N., Ceoldo, S., Tibaldo, M. & Sfriso, A. Trophic conditions in the waters of the Venice lagoon (Northern Adriatic Sea, Italy). *Open Oceanogr. J.* **5**, 1–13 (2011).
- Geertz-Hansen, O., Sand-Jensen, K., Hansen, D. F. & Christiansen, A. Growth and grazing control of abundance of the marine macroalga, *Ulva lactuca* L. in a eutrophic Danish estuary. *Aquat. Bot.* **46**, 101–109 (1993).
- Kamermans, P. et al. Effect of grazing by isopods and amphipods on growth of *Ulva* spp. (*Chlorophyta*). *Aquat. Ecol.* **36**, 425–433 (2002).
- Bäck, S., Lehvo, A. & Blomster, J. Mass occurrence of unattached *Enteromorpha intestinalis* on the Finnish Baltic Sea coast. *Ann. Bot. Fenn.* **37**, 155–161 (2000).
- Lin, H. Z. et al. Genetic and marine cyclonic eddy analyses on the largest macroalgal bloom in the world. *Environ. Sci. Technol.* **45**, 5996–6002 (2011).
- Zhang, X. W. et al. Somatic cells serve as a potential propagule bank of *Enteromorpha prolifera* forming a green tide in the Yellow Sea, China. *J. Appl. Phycol.* **22**, 173–180 (2010).
- Zhang, J. H. et al. Growth characteristics and reproductive capability of green tide algae in Rudong coast, China. *J. Appl. Phycol.* **25**, 795–803 (2013).
- Viscusi, G. Fear of noxious 'green tides' drives tourists from beaches of Brittany http://www.boston.com/news/science/articles/2011/08/14/fear_of_noxious_green_tides_drives_tourists_from_beaches_of_brittany/ (Boston Globe, 2011).
- Diaz, M., Darnhofer, I., Darrot, C. & Beuret, J.-E. Green tides in Brittany: What can we learn about niche–regime interactions? *Environ. Innov. Soc. Transitions* **8**, 62–75 (2013).
- Samuel, H. French protesters say Brittany will be François Hollande's 'cemetery' <http://www.telegraph.co.uk/news/worldnews/francois-hollande/10423888/French-protesters-say-Brittany-will-be-Francois-Hollandes-cemetery.html> (The Telegraph, 2013).
- Jing, L. Seaweed farming linked to Qingdao's green tide of algae <http://www.scmp.com/news/china/article/1284156/cause-qingdaos-green-tide-algae-mystery> (South China Morning Post, 2013).
- Williams, A. & Feagin, R. *Sargassum* as a natural solution to enhance dune plant growth. *Environ. Manage.* **46**, 738–747 (2010).
- Gower, J. & King, S. Distribution of floating *Sargassum* in the Gulf of Mexico and the Atlantic Ocean mapped using MERIS. *Int. J. Remote Sens.* **32**, 1917–1929 (2011).
- Johnson, D. L. & Richardson, P. L. On the wind-induced sinking of *Sargassum*. *J. Exp. Mar. Biol. Ecol.* **28**, 255–267 (1977).
- Hemphill, A. Change is in the air – seaweed, seaweed everywhere! <http://arlohemphill.com/2011/08/26/change-is-in-the-air-seaweed-seaweed-everywhere/>. (Arlo Hemphill, 2013).
- Gower, J., Young, E. & King, S. Satellite images suggest a new *Sargassum* source region in 2011. *Remote Sens. Lett.* **4**, 764–773 (2013).
- Ackah-Baidoo, A. Fishing in troubled waters: oil production, seaweed and community-level grievances in the Western Region of Ghana. *Community Dev. J.* **48**, 406–420 (2013).
- McDiarmid, J. Western Ghana's fisherfolk starve amid algae infestation <http://www.ipsnews.net/2012/04/western-ghanas-fisherfolk-starve-amid-algae-infestation/> (IPS, 2011).
- Froese, R. & Pauly, D. (eds). FishBase. <http://www.fishbase.org> (Fishbase, 2013).
- Anderson, D. M., Cembella, A. D. & Hallegraeff, G. M. Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annu. Rev. Mar. Sci.* **4**, 143–176 (2012).
- GEOHAB. *Global Ecology and Oceanography of Harmful Algal Blooms* (SCOR and IOC, 2001).
- Nelson, T. A., Lee, D. J. & Smith, B. C. Are "green tides" harmful algal blooms? Toxic properties of water-soluble extracts from two bloom-forming macroalgae, *Ulva fenestrata* and *Ulva obscura* (*Ulvoephyceae*). *J. Phycol.* **39**, 874–879 (2003).
- Harder, T., Dobretsov, S. & Qian, P.-Y. Waterborne polar macromolecules act as algal antifoulants in the seaweed *Ulva reticulata*. *Mar. Ecol. Prog. Ser.* **274**, 133–141 (2004).
- Algae Industry Magazine. *Olmix opens algae biorefinery in Brittany* <http://www.algaeindustrymagazine.com/olmix-opens-algae-biorefinery-brittany/> (Algae Industry Magazine, 2013).
- South Atlantic Fishery Management Council. *Fishery Management Plan For Pelagic Sargassum Habitat Of The South Atlantic Region* http://www.gc.noaa.gov/documents/gcil_safmc_fmp.pdf (NOAA, 2002).
- Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. & Torres, F. Fishing down marine food webs. *Science* **279**, 860–863 (1998).

Acknowledgments We thank C. Barroca, A. Huckbody, E. Fuller and M. Freling for sharing their photographs and experience, I. Valiela for comments on an earlier draft and P. Kullberg for updates.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/cekqww. Correspondence should be addressed to V.S. (victor.smetacek@awi.de) or A.Z. (zingone@szn.it).

FORUM: Planetary science

Shadows cast on Moon's origin

Our knowledge of how Earth's natural satellite formed is increasingly being challenged by observations and computer simulations. Two scientists outline our current understanding from the point of view of the satellite's geochemistry and its early dynamical history.

A chip off the old block

TIM ELLIOTT

Not since NASA's scientists definitively announced that the lunar white stuff was non-dairy has the Moon faced such an identity crisis. Ironically, it seems that our satellite is compositionally too similar to Earth for a simple explanation of its origins. Most dynamical and even some chemical attributes of the Earth–Moon system have been successfully explained by a 'giant impact' scenario, in which a Mars-sized impactor collided with the proto-Earth. Yet the standard version of this model produces a Moon that is mainly made of the impactor and not the target (Fig. 1). As emphasized in a Royal Society meeting¹ in September that debated the origin of the Moon, the compositional differences between Earth and the Moon that would be expected as a consequence are increasingly at odds with diverse, high-precision isotopic observations.

The isotopic kinship of Earth and the Moon was initially apparent in their indistinguishable oxygen isotope ratios², which contrasted with analyses of meteorite samples from most other planetary objects in the Solar System. The dilemma of this matched isotopic composition has deepened with more-recent measurements — notably, analyses of tungsten³ and silicon⁴ isotopes, which are controlled by very different processes from oxygen.

The radiogenic-tungsten isotope ratios of different planetary mantles should vary because they record the stochastic growth of the parent bodies and the formation of their cores. For the impactor and the proto-Earth to have the same oxygen isotope ratio is unlikely², but for them also to have the same tungsten isotopic composition is highly implausible. The distinctive silicon isotopic composition of Earth's silicate mantle reflects the consequences of silicon sequestration by a core formed at high temperatures on a large planetary body. Despite its moniker, the Mars-sized impactor of the standard giant-impact

model is not large enough to have experienced conditions that would generate an Earth-like silicon isotope ratio. Thus, differences in oxygen, tungsten and silicon isotope ratios between target and impactor seem inevitable, and so the standard model predicts isotopic differences between Earth and the Moon that are not observed.

These various isotopic embarrassments might potentially be explained away by rapid isotopic re-equilibration of Earth and the Moon in the vapour-rich aftermath of the Moon-forming collision⁵. But recent work has shown⁶ that the isotopic similarity of the two bodies extends to refractory elements such as titanium, which should not remain in the vapour phase long enough to allow such re-equilibration.

New dynamical models that can produce the Moon from the proto-Earth do not have the inherent simplicity of the canonical giant-impact scenario, and some argue that there are crucial flaws in such models. The sequence of conditions that currently seems necessary in these revised versions of lunar formation have led to philosophical disquiet. From a naive geochemical perspective, however, the isotopic similarity of Earth and the Moon holds an obvious appeal; the proto-Earth represents an abundant local source of material from which to build the Moon. Whether or not this comfort of availability can be meshed with the rigours of celestial mechanics remains to be seen.

Tim Elliott is in the School of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK. e-mail: tim.elliott@bristol.ac.uk

Weak links mar lunar model

SARAH T. STEWART

The giant-impact hypothesis of lunar origin is celebrated for its simplicity: a late, grazing impact on the proto-Earth launches a portion of the rocky mantle into orbit and

establishes the angular momentum of the Earth–Moon system (Fig. 1b). The Moon, depleted of iron and volatile elements relative to Earth, forms from this hot circumterrestrial disk of rocky mantle. Hydrodynamic simulations of giant impacts successfully produce disks of low iron content and sufficient mass to make this hypothesis plausible. The fatal issue is that simulations that lead to the present angular momentum derive most disk material from the impactor⁷. Thus, the giant-impact model predicts that Earth and the Moon should be derived from different source material, each with distinct isotopic fingerprints, and this contradicts the geochemical (isotopic) observations.

A possible way forward relaxes the constraints on the giant-impact model. Perhaps it was too much to ask that a single dynamical process should satisfy all the physical and geochemical observations. In fact, formation of the Earth–Moon system is thought to have been an extended, multi-stage process: a giant impact creates a disk (on a timescale of 1 day), the Moon accretes from the disk (hundreds of years), and interactions known as orbital resonances, which occur during lunar tidal evolution, establish the inclination and angular momentum of the system (up to tens of thousands of years).

However, two studies^{8,9} last year proposed different giant-impact scenarios for generating a disk that is compositionally similar to Earth, and so meet the isotopic observations (Fig. 1c, d). These leave Earth spinning near the limit of its stability and require a separate mechanism by which the Earth–Moon system reaches the present-day angular momentum. The evection resonance, which occurs when the short axis of the Moon's elliptical orbit about Earth rotates synchronously with the orbit of Earth around the Sun, is encountered quickly during the tidal evolution of the Moon and could have transferred the excess angular momentum away from the Earth–Moon system⁸. These new

solutions have broken the stalemate between the models and the geochemical data, and

NATURE.COM
For more on the Moon's origins, see: go.nature.com/5foh6i

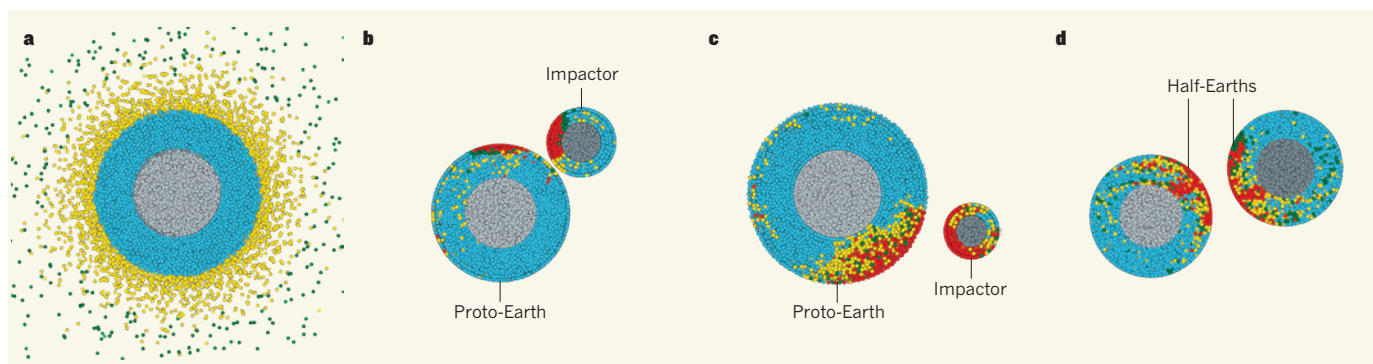


Figure 1 | Making Earth and the Moon into isotopic twins. After the giant impact, the Moon forms from a disk of material around Earth. In these views of an equatorial slice through the post-impact Earth (a) and impacting bodies (b–d), colours denote material that ends up in the core (grey), mantle (cyan), hot silicate atmosphere (yellow) and lunar disk (green). Red material escapes the Earth–Moon system. In the canonical giant-impact model⁷, the lunar material is derived primarily from the impactor's mantle and the shallow mantle of the proto-Earth (b). Material from these sources is not expected to be identical to the bulk silicate Earth (see, for example, refs 2–4). In the new giant-impact models^{8,9}, lunar material is derived either from a range of depths in the proto-Earth's mantle (c) or equally from the entire mantles of two colliding half-Earths (d). These sources are more likely to produce a Moon with the same isotopic fingerprint as Earth.

have shifted attention towards the weak links between the major stages of lunar origin.

Now, the lunar origin cannot be addressed by a single (rather simple) hydrodynamic calculation. Modelling the formation of the Moon in greater detail poses challenges to both our understanding of the physics of what occurred and our technical capabilities. For example, seeding the lunar disk with Earth-mantle material might not be sufficient to explain the isotopic similarity. The initial conditions of the disk have not been robustly established by the hydrodynamic calculations, which neglect the disk's many chemical components and its multiphase flow. Coupled dynamical and chemical models for the lunar disk are in their infancy; mixing within the disk might eliminate some of the initial

differences or generate new chemical differences during its evolution. Crucially, the time during which the Moon is caught in the evection resonance, the crux of the new class of impact scenario, is sensitive to its thermal state¹, which depends on the details of lunar accretion from the disk.

Within our current understanding of planetary and satellite formation processes, each stage of lunar evolution is plausible. But, with the nested levels of dependency in a multi-stage model, is the probability of the required sequence of events vanishingly small? Is there an alternative solution of greater simplicity and universality? Ultimately, the current detailed interrogation of lunar origin may demand answers that have an unexpected level of complexity. ■ [SEE COMMENT P.27](#)

Sarah T. Stewart is in the Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mail: ss Stewart@eps.harvard.edu

1. <http://royalsociety.org/events/2013/moon-origin-satellite>
2. Clayton, R. N. & Mayeda, T. K. *Geochim. Cosmochim. Acta* **60**, 1999–2017 (1996).
3. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. *Nature* **450**, 1206–1209 (2007).
4. Armytage, R. M. G., Georg, R. B., Williams, H. M. & Halliday, A. N. *Geochim. Cosmochim. Acta* **77**, 504–514 (2012).
5. Pahlavan, K. & Stevenson, D. J. *Earth Planet. Sci. Lett.* **262**, 438–449 (2007).
6. Zhang, J., Dauphas, N., Davis, A. M., Leya, I. & Fedkin, A. *Nature Geosci.* **5**, 251–255 (2012).
7. Canup, R. M. *Icarus* **196**, 518–538 (2008).
8. Čuk, M. & Stewart, S. T. *Science* **338**, 1047–1052 (2012).
9. Canup, R. M. *Science* **338**, 1052–1055 (2012).

STEM CELLS

Dual response to Ras mutation

Proliferation-driving mutations in haematopoietic stem cells often result in the loss of stem-cell properties. But at least one common oncogenic mutation seems to enhance both proliferation and stem-cell self-renewal. [SEE LETTER P.143](#)

S. HAIHUA CHU & SCOTT A. ARMSTRONG

When a stem cell divides, it can either produce differentiated cells or self-renew to produce more stem cells. Because stem cells are thought to be the cells of origin for many types of cancer, understanding what controls this decision has become a central question in stem-cell and cancer research. During the formation of mature blood cells from haematopoietic stem cells (HSCs), these processes are often diametrically opposed: blood

cells are produced through a hierarchical process of proliferation and differentiation, often at the expense of the stem-cell ability to self-renew. But how is this decision altered when a stem cell acquires a cancer-driving mutation? Previous studies have shown that mutations that increase the proliferation of HSCs tend to reduce the cells' potential for self-renewal. But on page 143 of this issue, Li *et al.*¹ report that HSCs harbouring an activating mutation of the protein *Nras* show not only enhanced proliferation but also enhanced self-renewal.

Nras is a member of the Ras family of proteins, which transmit cellular proliferation and survival signals in many different contexts and which are frequently mutated to become constitutively active in cancer cells. Li and colleagues found in mice that expression of an activating mutant version of the *Nras* gene in HSCs led to an increased number of the cells entering the cell cycle. In line with previous observations², the *Nras*-mutant HSCs outcompeted normal HSCs in their ability to reconstitute haematopoiesis when both cell types were transplanted into HSC-depleted mice. But surprisingly, the researchers also found that *Nras*-mutant HSCs could be serially transplanted in mice through more rounds of transplantation than normal cells, demonstrating enhanced self-renewal.

To determine how one signalling molecule, mutant *Nras*, could confer both enhanced proliferation and self-renewal potential on HSCs, Li *et al.* used *Nras*-expressing HSCs that expressed a fluorescent 'reporter' protein, so that they could monitor cell division by the dilution of fluorescence over time. Remarkably, they observed two distinct responses: mutant *Nras* reduced the division and increased the

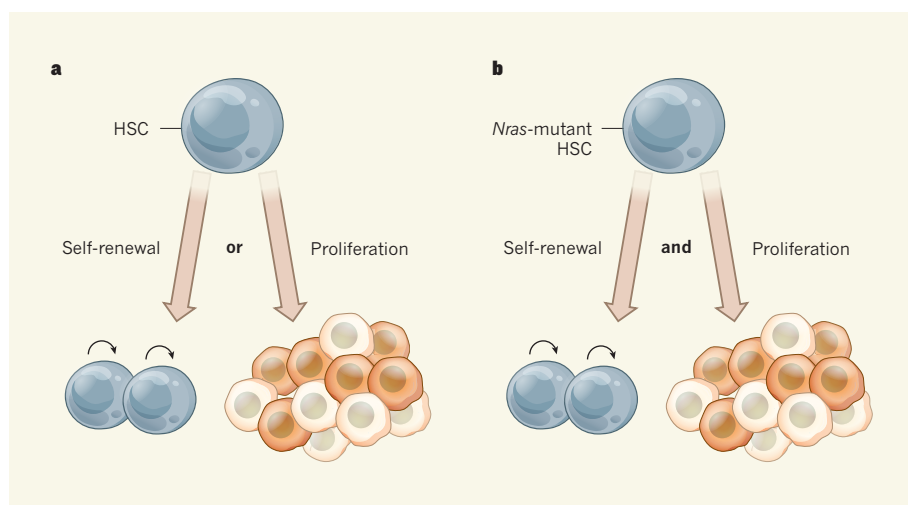


Figure 1 | Bimodal behaviour. **a**, Division of a haematopoietic stem cell (HSC) typically results in either proliferation of more-differentiated cells or self-renewal of the HSC. **b**, HSCs harbouring an activating mutation in *Nras*, however, show both enhanced proliferation and self-renewal¹, possibly because *Nras* activation has different effects on different HSC subsets.

self-renewal potential of one subset of HSCs, but increased the division and reduced the self-renewal potential of another subset. These findings suggest that there is a bimodal response to *Nras* activation in HSCs (Fig. 1).

The authors then studied signalling pathways downstream of *Nras* in HSCs, and observed not only activation of the MEK–ERK kinase pathway as expected, but also activation of the STAT5 signalling pathway, which is not well known as an effector of activated Ras proteins. Remarkably, deletion of just one of the two copies of the gene encoding STAT5 in the *Nras*-mutant HSCs attenuated the increase in both proliferation and self-renewal, suggesting that STAT5 could be a therapeutic target that eradicates not only the rapidly proliferating subset of cells, but also those cells that have enhanced self-renewal and are therefore more quiescent. These findings are of particular interest because STAT5 has been previously implicated in Ras-driven haematopoietic malignancies³, and MEK inhibition alone had a variable effect on cycling HSCs in these studies. In general, MEK inactivation does not reliably eliminate mutant HSCs in mouse models of Ras-pathway activation⁴. But Li and colleagues' findings indicate that mutant *Nras* induces aberrant signalling in HSCs that could be exploited therapeutically.

There is an expanding body of literature suggesting heterogeneity and diversity of function in HSCs. Of the most primitive (least differentiated) HSCs, some are poised for proliferation and differentiation, whereas others are programmed for quiescence, and these cell populations exist in a dynamic equilibrium⁵. It is also known that individual HSCs do not have identical lineage potential^{6,7}. The bimodal effect defined by Li *et al.* could be explained by this functional heterogeneity, with some cells responding to mutation by enhancing their self-renewal potential and quiescence, and others responding with enhanced proliferation

and differentiation. This leads to the question of whether such bimodal behaviour is the result of a stochastic response or is determined by definable subsets of HSC responding in a predictable fashion. In light of recent data from normal HSCs^{5–7}, we would predict the latter, but further work is required to test this.

Another intriguing question raised by this study is whether the bimodal effect of activating mutations is unique to *Nras*, or whether it is also a feature of other oncogenic mutations, such as the mutations in the Ras family member *Kras* that are seen in many solid tumours and in some haematopoietic cancers. In mouse models, both *Kras*⁸ and *Nras*⁹ mutations lead to excess production of cells from the bone marrow that seems to be initiated from primitive HSCs, although the type of leukaemia that arises differs (T-ALL and AML, respectively).

It is unclear whether *Kras* activation increases HSC self-renewal potential in a similar manner to *Nras* activation, and this prompts the question of whether HSC responses to different Ras mutations could determine the type of leukaemia that develops. It will be important to study not only the similarities and differences in how these and other oncogenic mutations affect cell signalling and gene expression, but also the cellular contexts required for such responses.

It is becoming increasingly clear that cellular context can influence the response to an oncogenic mutation^{10,11}. This suggests that it is no longer sufficient to know simply which mutations are present in a tumour; we must also consider the influence of where and when. Thus, as we learn more about the mutations that occur in tumour cells, we will need to assess these mutations in functional assays such as those described by Li *et al.*, to obtain a more accurate picture of the effects of mutations in specific cell types and at specific points in development. Such experiments will further enhance our understanding of the complex cellular heterogeneity found in cancers. ■

S. Haihua Chu and Scott A. Armstrong are at the Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. e-mail: armstros@mskcc.org

1. Li, Q. *et al.* *Nature* **504**, 143–147 (2013).
2. Wang, J. *et al.* *Blood* **121**, 5203–5207 (2013).
3. Kotecha, N. *et al.* *Cancer Cell* **14**, 335–343 (2008).
4. Chang, T. *et al.* *J. Clin. Invest.* **123**, 335–339 (2013).
5. Wilson, A. *et al.* *Cell* **135**, 1118–1129 (2008).
6. Dykstra, B. *et al.* *Cell Stem Cell* **1**, 218–229 (2007).
7. Yamamoto, R. *et al.* *Cell* **154**, 1112–1126 (2013).
8. Sabnis, A. J. *et al.* *PLoS Biol.* **7**, e1000059 (2009).
9. Li, Q. *et al.* *Blood* **117**, 2022–2032 (2011).
10. Wang, Y. *et al.* *Science* **327**, 1650–1653 (2010).
11. Friedmann-Morvinski, D. *et al.* *Science* **338**, 1080–1084 (2012).

This article was published online on 27 November 2013.

ASTROPHYSICS

Magnetic fields in γ -ray bursts

Observations of a high degree of polarization in the immediate optical afterglow of a γ -ray burst indicate that these powerful cosmic explosions carry large-scale, ordered magnetic fields. [SEE LETTER P.119](#)

MAXIM LYUTIKOV

Naturally occurring magnetic fields protect life on Earth from energetic cosmic rays but are relatively weak and barely noticeable. However, for astronomical objects, magnetic fields can have a dynamically important, and often dominant, role, especially

for gravitationally collapsed objects such as neutron stars and accreting black holes. On page 119 of this issue, Mundell *et al.*¹ report the possible observation of an ordered magnetic field that plays a significant part in a cosmic explosion known as a γ -ray burst.

As a black hole gravitationally pulls matter in, magnetic fields, which are frozen into the

accreting plasma owing to the plasma's high conductivity, are compressed and thus amplified. Compressed magnetic fields can produce spectacular astrophysical phenomena because of a key ingredient: rotation. Collapsed objects, as well as the material that they accrete, typically rotate rapidly. The combination of rotation and compressed magnetic fields leads to the astronomical realization of the Faraday wheel — an electric generator of constant electrical polarity — that produces large currents and voltages².

This is how powerful astrophysical outflows such as γ -ray bursts (GRBs) are produced: material accreting onto a rotating black hole brings with it the magnetic field and sets up a mega-version of the Faraday wheel^{3,4}. The resulting plasma outflow may reach extremely high (relativistic) velocities and, in the case of minutes-long GRBs, carry energy comparable to the total energy that the Sun will radiate over its multibillion-year lifetime⁵. In addition, hoop stresses produced by the large-scale helical magnetic field that permeates the plasma outflow can collimate the outflow into a narrow beam, with an opening angle of only a few degrees.

As a GRB outflow interacts with the surrounding medium, two shocks are launched: a forward shock into the external medium and a reverse shock into the ejecta. Observations of the emission from the reverse shock, whose frequency typically falls in the optical range, can probe the properties of the ejecta and the Faraday-wheel model. Testing this model requires verification that the outflow carries a large-scale, ordered magnetic field (assuming that the field extends well into the outflow).

In their study, Mundell *et al.* report the possible detection of just such a magnetic field, in a GRB dubbed GRB 120308A, through observations of polarized, early optical emission from a reverse shock in the GRB. Medium-sized optical telescopes, such as the 2-metre Liverpool Telescope used in the present study, can slew to the position of the burst within minutes of receiving the alert that the burst has occurred, and detect a typically faint optical afterglow. But in the case reported here, the optical telescope was also equipped with a purpose-built polarimeter, called RINGO2, that could detect the preferred orientation, or polarization, of the afterglow's oscillating electric field. Polarization indicates that the process that produced the afterglow is sensitive to a particular direction in the emitting plasma outflow.

Mundell and colleagues measured a linear polarization content — how much the electric field vibrates in a fixed plane — of 28% for the optical emission, with the angle of polarization remaining stable. This high value presumably comes from the reverse shock in the ejecta. It is close to the maximum degree of polarization that can be

produced by synchrotron-radiation-emitting electrons in a relativistically expanding outflow carrying a large-scale, ordered magnetic field.

This result is likely to contribute to the heated debate on the nature of GRBs. Despite decades of intensive research, we are still not clear about the basic, high-energy emission mechanism in GRBs, with the competing models being synchrotron emission from electrons and Compton scattering of photons by electrons. Previous claims⁶ of high polarization values in GRBs, which were based on observations made in the γ -ray energy regime, were inconclusive⁷ because polarization measurements are difficult to perform at high energies and subject to large uncertainties. By contrast, optical polarization, such as that obtained by the authors, can be

“This result is likely to contribute to the heated debate on the nature of γ -ray bursts.”

STRUCTURAL BIOLOGY

Ion channel seen by electron microscopy

Structures of the heat-sensitive TRPV1 ion channel have been solved using single-particle electron cryo-microscopy, representing a landmark in the use of this technique for structural biology. SEE ARTICLES P.107 & P.113

RICHARD HENDERSON

Membrane proteins known as transient receptor potential (TRP) ion channels occur in species ranging from yeast to humans. Members of this receptor family are involved in the perception of an enormous range of stimuli¹, including vision (in invertebrates), taste, hot or cold temperatures, pH and physical forces. On page 107 of this issue, Liao *et al.*² report the first high-resolution structure of TRPV1, the ion channel responsible for sensing heat. And in a second paper from the same group, Cao *et al.*³ (page 113) describe the sites at which three ligand molecules bind to TRPV1, and how this binding triggers the opening of the channel.

There are 27 members of the TRP receptor family in humans, each with their own functions and different tissue distribution. Most TRP channels, including TRPV1, are weakly selective for calcium ions. TRPV1 was first identified⁴ as the receptor for capsaicin — the compound that makes chilli peppers seem hot — in 1997. The channel has four identical subunits, and the modified version used in the

measured with much higher certainty. Mundell and colleagues' detection of a high degree of polarization in the optical afterglow both confirms the Faraday-wheel model of launching powerful astrophysical outflows and argues in favour of synchrotron radiation being the dominant high-energy emission process. ■

Maxim Lyutikov is in the Department of Physics, Purdue University, West Lafayette, Indiana 47907-2036, USA.
e-mail: lyutikov@purdue.edu

1. Mundell, C. G. *et al.* *Nature* **504**, 119–121 (2013).
2. Blandford, R. D. & Znajek, R. L. *Mon. Not. R. Astron. Soc.* **179**, 433–456 (1977).
3. Blandford, R. D. in *Lighthouses of the Universe: The Most Luminous Celestial Objects and Their Use for Cosmology* (eds Gilfanov, M., Sunyaev, R. & Churazov, E.) 381–404 (Springer, 2002).
4. Lyutikov, M. *New J. Phys.* **8**, 119 (2006).
5. Gehrels, N., Ramirez-Ruiz, E. & Fox, D. B. *Annu. Rev. Astron. Astrophys.* **47**, 567–617 (2009).
6. Coburn, W. & Boggs, S. E. *Nature* **423**, 415–417 (2003).
7. Rutledge, R. E. & Fox, D. B. *Mon. Not. R. Astron. Soc.* **350**, 1288–1300 (2004).

present studies has an overall molecular weight of about 300 kilodaltons (bigger than most ion channels). Not only is TRPV1 opened by capsaicin, it is also strongly activated by toxins, such as resiniferatoxin from *Euphorbia* plants, or 'cysteine-knot' toxins from tarantulas. These chemosensory stimuli are thought to have evolved as protective deterrents against predators, and elicit a burning sensation by usurping normal heat sensing through TRPV1 activation.

To solve the structure of TRPV1, Liao *et al.* used single-particle electron cryo-microscopy (cryo-EM), with no help from any of the more established methods of structural biology. The authors made full use of several technical advances: they used a slightly truncated rat TRPV1 construct that is biochemically stable; they transferred purified ion channels into a polymeric 'amphipol' framework⁵ to maintain the channels' stability and solubility in water; and, most importantly, they used a camera that detects electrons directly (minimizing noise and allowing any image blurring during an exposure to be compensated for^{6,7}) and a state-of-the-art computer program that

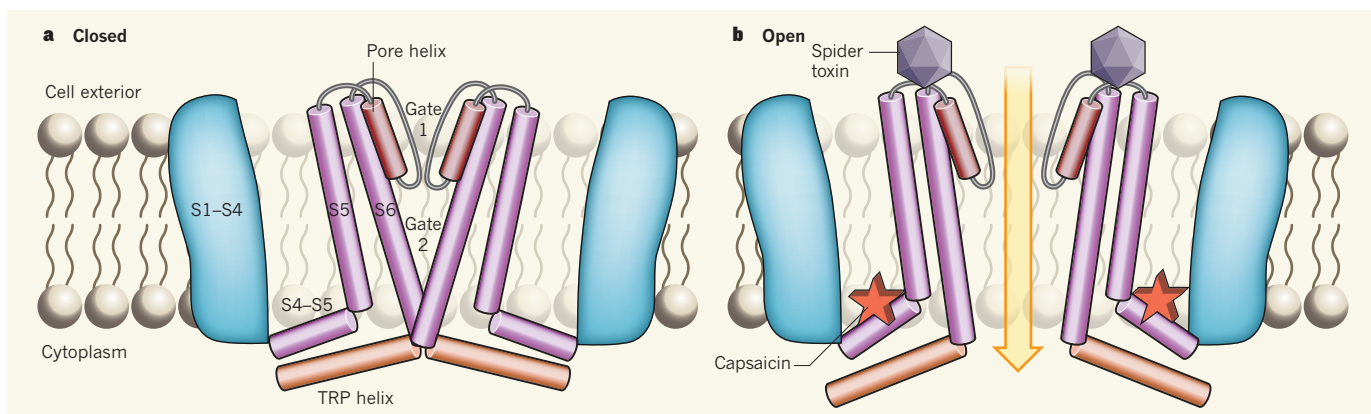


Figure 1 | TRPV1 in closed and open states. The cartoon shows cross sections of the closed and open states of the TRPV1 ion channel, based on structures^{2,3} obtained using electron cryo-microscopy. The channels form from four subunits, only two of which are shown. Helices in one subunit (S5, S6, the S4–S5 linker, the pore helix and the TRP helix) are labelled. A bundle of helices S1 to S4 is depicted as a single object. **a**, Two gates are apparent in the closed state, one near the extracellular surface (gate 1) and another deeper within the channel (gate 2). **b**, Gate 1 opens in response to the binding of spider toxin, whereas gate 2 opens on binding of capsaicin. Another ligand, resiniferatoxin (not shown), can bind at the same site as capsaicin. The arrow indicates the passage of calcium ions through the channel.

uses statistical methods to estimate particle orientations and to calculate highly accurate reconstructions of structures⁸.

Liao and colleagues' analysis advanced from an initial EM structure obtained at a resolution of 30 Å, through a cryo-EM structure at 8.8 Å resolution, to a final map that includes regions depicted at a resolution of 3.4 Å — good enough for amino-acid side chains and β -sheets to be recognized, and for the polypeptide backbone of the protein to be traced. Previous work^{9,10} in which cryo-EM was used to study TRP channels reached resolutions of only 15–19 Å. The current work is therefore a landmark both in the evolution of the single-particle cryo-EM method and in its use for tackling the structure of a macromolecular complex that has been difficult to study by other means. Notably, the regions of the map with highest resolution are near the centre of the molecule. More work is needed to discover whether the lower resolution in peripheral regions is caused by the limited accuracy with which orientations of single protein particles can be determined or by flexibility of the TRPV1 channel in these regions.

The authors confirm that the overall architecture of TRP channels is similar to that of members of another group of ion channels, the voltage-gated sodium and potassium channels¹¹. The members of this superfamily all contain a central ion-conducting pore that is made up of two carboxy-terminal transmembrane α -helices (S5 and S6) and a loop containing a short helix (the pore helix) from each subunit (or from each domain, in the case of sodium channels), together with a voltage-sensor module consisting of a bundle of four transmembrane helices (S1 to S4) at the amino terminus (Fig. 1). However, S4 in TRP channels contains only one or two of the characteristic basic amino-acid residues that are responsible for voltage gating (opening of the channel by a potential difference across the membrane). In this respect, TRP channels are

similar to a nucleotide-gated channel called MlotiK1 (ref. 12) because it has a compact, four-helix bundle that contains few basic residues, is less sensitive to transmembrane potentials and, as Cao *et al.* show, does not move during activation. Instead, high temperature or ligand binding triggers channel opening by a mechanism that is less dependent of voltage. But how?

Cao *et al.* find that the binding site for a large spider toxin lies distant from that for the small ligands resiniferatoxin and capsaicin. The spider toxin binds to the extracellular surface of the TRPV1 channel near the pore helix, with each of the toxin's two cysteine-knot domains binding to the junction between TRPV1 subunits. This locks open the extracellular end of the channel (gate 1 in Fig. 1).

The authors propose a binding site for capsaicin that is essentially the same as that for resiniferatoxin, although their identification of this site is less certain. The site consists of a cavity deep within the membrane towards the cytoplasmic side, surrounded by S3, S4, a linker between S4 and S5, and S6 from the adjacent subunit. The bound ligands can induce a structural change to the TRPV1 channel through close interactions with the S4–S5 linker, increasing the pore diameter by shifting S6 (gate 2 in Fig. 1). Thus, the TRPV1 channel seems to have two gates, one at either end of the channel. This dual gating, and the possible complexity of interactions between the two gates in response to different stimuli, is a principal finding of the authors' analysis.

An amino-acid motif adjacent to the C terminus of S6 contains the TRP box that is characteristic of TRP channels. Liao and colleagues' structure shows that the box consists of a short α -helix parallel to the membrane; this helix interacts with both the S4–S5 linker and another helix (the pre-S1 helix), and becomes less well ordered in the activated state.

The current work does not explain how TRPV1 is activated in response to temperature,

but this probably depends on finely balanced energy differences between its open and closed structures, which are difficult to address through structural analysis. Nevertheless, the availability of the new structures will surely help simulations of heat activation to be performed. The authors also hint that the methods used in their studies are well suited to trapping different states during heat-evoked gating.

Further improvements in detector efficiency, specimen-preparation methods and image-processing software should bring other advances in the use of cryo-EM for structural biology. With the outstanding success of the current work, the way is open for structural studies of many similar channels. Because TRPV1 and some other ion channels are potential targets for the development of pain-killing drugs, the findings may even herald the dawn of cryo-EM as a technique to aid rational drug design. ■

Richard Henderson is at the MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK.
e-mail: rh15@mrc-lmb.com.ac.uk

1. Venkatachalam, K. & Montell, C. *Annu. Rev. Biochem.* **76**, 387–417 (2007).
2. Liao, M., Cao, E., Julius, D. & Cheng, Y. *Nature* **504**, 107–112 (2013).
3. Cao, E., Liao, M., Cheng, Y. & Julius, D. *Nature* **504**, 113–118 (2013).
4. Caterina, M. J. *et al.* *Nature* **389**, 816–824 (1997).
5. Popot, J.-L. *et al.* *Annu. Rev. Biophys.* **40**, 379–408 (2011).
6. Li, X. *et al.* *Nature Meth.* **10**, 584–590 (2013).
7. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. *eLife* **2**, e00461 (2013).
8. Scheres, S. H. W. *J. Struct. Biol.* **180**, 519–530 (2012).
9. Mio, K. *et al.* *J. Mol. Biol.* **367**, 373–383 (2007).
10. Moiseenkova-Bell, V. Y., Stanciu, L. A., Serysheva, I. I., Tobe, B. J. & Wensel, T. G. *Proc. Natl Acad. Sci. USA* **105**, 7451–7455 (2008).
11. Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. *Nature* **450**, 376–382 (2007).
12. Clayton, G. M. *et al.* *Proc. Natl Acad. Sci. USA* **105**, 1511–1515 (2008).

NOBEL 2013

As the recipients of the 2013 science Nobel prizes gather in Stockholm to celebrate and be celebrated, News & Views shares some expert opinions on the achievements honoured.

CHEMISTRY

Methods for computational chemistry

The Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for their work on developing multiscale models for complex chemical systems (see figure).

MULTISCALE MODELS by Walter Thiel

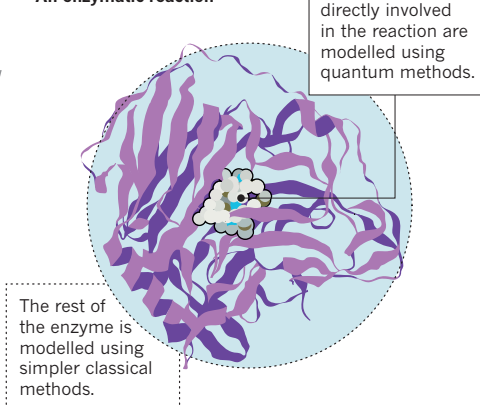
Complex chemical processes occur on different length- and timescales. Events that involve electrons, such as the making and breaking of chemical bonds, are localized in space and time. They need to be described by quantum mechanics (QM), whereas the influence of the environment and the slow motions of atoms during a reaction are normally well represented by classical molecular mechanics (MM).

The laureates were the first to propose a hybrid QM/MM approach for studying chemical properties and reactions, initially for the special case of planar molecules¹ and then as a general scheme for modelling enzymatic reactions². This mathematical approach is essentially a marriage of Schrödinger's quantum theories and classical Newtonian ideas, combining the best of both worlds to enable tailor-made simulations of complex chemical processes.

The prizewinners' pioneering work in the 1970s provided explicit expressions for calculating the total QM/MM energy of a system and the QM/MM interaction terms. Advances by many research groups in QM and MM methods during the 1980s paved the way to breakthroughs for QM/MM modelling in chemistry in the 1990s. Major methodological issues were then solved by establishing

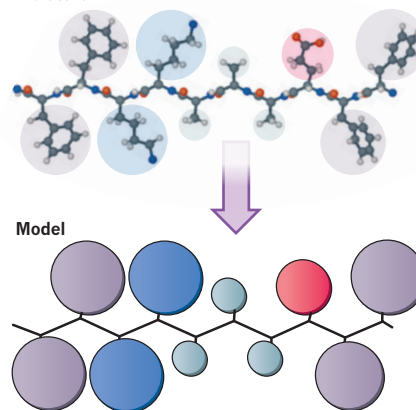
Karplus, Levitt and Warshel married classical and quantum methods to model complex chemical processes computationally^{1,2}.

An enzymatic reaction



Warshel and Levitt also showed that groups of atoms can be treated as rigid units to speed up modelling of large systems.

Molecule



suitable QM/MM interaction models and treatments for the QM/MM interface region, and efficient procedures were implemented for exploring large-scale QM/MM potential surfaces (which represent total energy as a function of atomic position). Since then, there has been an exponential growth in QM/MM applications³, all underpinned by the original work of Karplus, Levitt and Warshel.

The concept of multiscale modelling is actually much broader than QM/MM, and so one can safely expect further progress towards an ever more realistic modelling of increasingly complex chemical processes.

COMPUTER EXPERIMENTS by Gerhard Hummer

Multiscale molecular simulations, as pioneered by Karplus, Levitt and Warshel, proved to be versatile and powerful right from the start, revealing how receptors in the eye are activated by light, and how the resulting signals are passed on through changes in molecular conformation.

The laureates' approach allows each part of a molecular system to be described at the simplest level possible: as atoms, using quantum

or classical mechanics; as classical pseudo-particles that represent multiple atoms; or, in the case of bulk solvent, as a continuous medium that lacks atomic detail^{2,4}. Molecular interactions are captured by potential surfaces. Such potentials are now used routinely to determine protein structures from experimental data, to develop new drugs and to rationally design materials.

Simulations also provide fundamental insight into the function of biomolecular 'machinery' by revealing the underlying molecular motions and energetic driving forces. From photosynthesis to the processing of genetic material⁵, enzyme-catalysed reactions have been modelled and followed atom by atom, bond by bond³. The dynamics of molecular motors that power muscle contraction or the synthesis of ATP molecules — a cell's source of energy — have also been simulated. Even the self-assembly of biomolecular machinery can be studied, from the folding of proteins⁶ to the formation of entire organelles⁷ and the protein shells of viruses⁸.

With increasingly accurate representations of the energetics and dynamics of molecular systems, simulations yield detailed quantitative information and mechanistic insight that are unattainable in laboratory experiments. The vision of computational modelling as a

reliable substitute for actual experiments is thus becoming a reality. ■

Walter Thiel is at the Max-Planck-Institut für Kohlenforschung, 45470 Mülheim an der Ruhr, Germany. **Gerhard Hummer** is at the Max-Planck-Institut für Biophysik,

60438 Frankfurt am Main, Germany.
e-mails: thiel@kofo.mpg.de;
gerhard.hummer@biophys.mpg.de

1. Warshel, A. & Karplus, M. *J. Am. Chem. Soc.* **94**, 5612–5625 (1972).
2. Warshel, A. & Levitt, M. *J. Mol. Biol.* **103**, 227–249 (1976).

3. Senn, H. M. & Thiel, W. *Angew. Chem. Int. Edn* **48**, 1198–1229 (2009).
4. Levitt, M. & Warshel, A. *Nature* **253**, 694–698 (1975).
5. Rosta, E., Nowotny, M., Yang, W. & Hummer, G. *J. Am. Chem. Soc.* **133**, 8934–8941 (2011).
6. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. *Science* **334**, 517–520 (2011).
7. Takamori, S. et al. *Cell* **127**, 831–846 (2006).
8. Zhao, G. et al. *Nature* **497**, 643–646 (2013).

ECONOMICS

Predicting asset prices

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was awarded to Eugene F. Fama, Lars Peter Hansen and Robert J. Shiller, whose empirical analysis of asset prices has shaped our understanding of how markets work (see figure).

EFFICIENCY AND VOLATILITY

by Christopher Polk

Fama's efficient market hypothesis (EMH) argues that competition among investors makes the return from using information on stock prices commensurate with the cost of that information. Thus, if costs are zero, prices correctly reflect all relevant information¹. According to this hypothesis, if we could easily predict that stock prices will rise tomorrow, we would all buy today, such that prices would in fact rise today until they reflected the information we had received. Tests by Fama in the 1960s found that short-run returns were mainly unpredictable, which is consistent with a market that incorporates information efficiently.

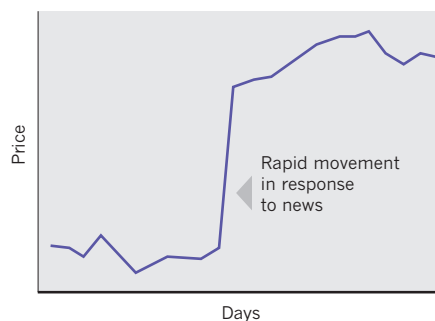
Fama emphasized that the EMH was not directly testable; one can only test a joint hypothesis of the EMH and a model detailing the way in which expected returns are set. If, say, small-company stocks generally outperform large-company stocks, this might not indicate that the pricing of small companies is inefficient, but rather that small-company stocks are riskier and hence their investors demand high expected returns as compensation².

In 1981, Shiller showed that historical prices were excessively volatile relative to their future realized value³. This suggested that although prices respond quickly to information, they change for other reasons as well. Shiller interpreted this volatility as resulting from investor sentiment. Subsequent work linked excess volatility to predictable variation in long-run returns; short-term predictability was later found as well.

These findings presented a serious challenge to the EMH, but Fama's joint hypothesis allows a possible explanation: time-varying

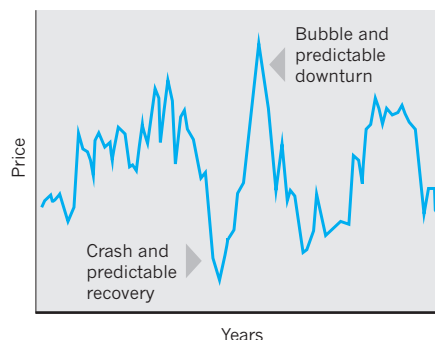
SHORT-TERM UNPREDICTABILITY

Fama showed that asset prices are extremely difficult to predict in the short term.



LONG-TERM PREDICTABILITY

But Shiller showed that there is greater predictability over the longer term, and interpreted this finding as market inefficiency resulting from investor behaviour.



TESTING THEORIES

Hansen's statistical techniques for testing economic theories highlighted the attractiveness of stocks to investors who can tolerate risk.



expected returns may be due to time-varying risk and/or risk aversion. Understanding the sources — rational and sentiment-based — of predictable variation in returns is at the heart of modern financial economics.

EMPIRICAL FINANCIAL ECONOMICS

by John Y. Campbell

Financial markets continually generate vast quantities of data on asset prices. Fama, Shiller and Hansen have led an effort, over almost 50 years, to use these data to better understand the economy and investor behaviour.

Fama observed that the return on any risky financial asset is the sum of a 'required' return that a rational investor expects to earn and an 'unexpected' return driven by the arrival of news. He noted that, over short time periods, the volatility of unexpected returns is much greater than any movement in the required return, and hence that short-term price movements accurately reflect the news hitting the market at each point of time.

Hansen built on Fama's insight, developing a powerful statistical method to extract from asset returns information about key properties of the economy, such as investors' average aversion to risk, without having to model other features of the economy that are irrelevant to the problem at hand^{4,5}.

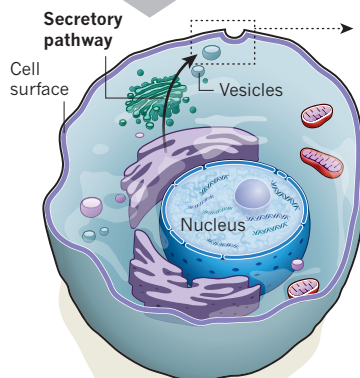
Shiller pointed to data indicating that large price swings result from the accumulation of movements in required returns over long periods of time, and that unexpected returns reflect not only news about the future payments that assets will make, but also unexpected changes in the required return⁶.

Together, their work has definitively shown the value of empirical research in understanding price formation in financial markets. Fama and Shiller have also used financial data to construct indexes that summarize the movements of broad categories of assets, such as groups of stocks with similar characteristics and houses in the same city². ■

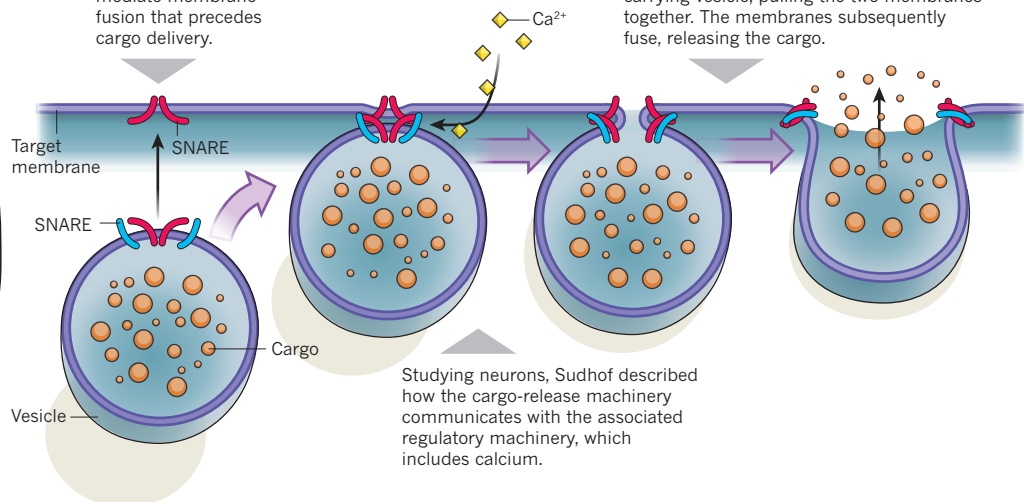
Christopher Polk is in the Department of Finance, London School of Economics, London WC2A 2AE, UK. **John Y. Campbell** is in the Department of Economics, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mails: c.polk@lse.ac.uk;
john_campbell@harvard.edu

1. Fama, E. F. *J. Finance* **25**, 383–417 (1970).
2. Fama, E. F. & French, K. R. *J. Financ. Econ.* **33**, 3–56 (1993).
3. Shiller, R. *J. Am. Econ. Rev.* **71**, 421–436 (1981).
4. Hansen, L. P. *Econometrica* **50**, 1029–1054 (1982).
5. Hansen, L. P. & Jagannathan, R. *J. Pol. Econ.* **99**, 225–262 (1991).
6. Campbell, J. Y. & Shiller, R. *J. Rev. Financ. Stud.* **1**, 195–228 (1988).

Schekman identified many of the genes that control intracellular transport of the cargo-carrying vesicles along the secretory pathway in yeast cells.



Rothman revealed that SNARE proteins mediate membrane fusion that precedes cargo delivery.



On the target membrane, SNARE proteins 'zip up' with specific SNAREs on the cargo-carrying vesicle, pulling the two membranes together. The membranes subsequently fuse, releasing the cargo.

Studying neurons, Südhof described how the cargo-release machinery communicates with the associated regulatory machinery, which includes calcium.

PHYSIOLOGY OR MEDICINE

Traffic control system within cells

The recipients of the Nobel Prize in Physiology or Medicine are Randy W. Schekman, James E. Rothman and Thomas C. Südhof, for their discoveries on how cells deliver thousands of internally generated molecules to the right place at the right time (see figure).

IDENTIFICATION OF THE TRAFFICKING MACHINERY

by Susan Ferro-Novick

In the late 1970s, Randy Schekman and Jim Rothman both strove to identify the cellular machinery that drives the secretory pathway, albeit by taking strikingly different approaches. Schekman and co-workers took advantage of yeast genetics to initially identify 23 genes, called *SEC*, whose products are required for protein secretion^{1,2}. Rothman and colleagues used a biochemical approach to purify, by brute force, components of the mammalian secretory apparatus³.

As the *SEC* genes were characterized and several of their mammalian counterparts were identified, it became clear that Schekman and Rothman's independent approaches had converged and catalysed the field of membrane traffic. Schekman went on to focus on the 'coat' proteins that sort cellular cargo into a nascent vesicle. Rothman instead focused on

the membrane-fusion machinery, which he called SNARE proteins⁴, and which is found in all eukaryotes (organisms that include fungi, plants and animals).

The ground-breaking work of these laureates has revolutionized our understanding of a basic cellular function — protein secretion. Whereas the genetic approach identified many of the components of the pathway, the biochemical assays facilitated elucidation of the components' functions. The laureates' seminal contributions paved the way for studies of many other cellular processes that rely on the secretory pathway, including cell polarization, cell migration and the degradative process of autophagy.

A TURBOCHARGER FOR MEMBRANE FUSION

by Nils Brose

The realization that the membrane-fusion machinery is evolutionarily conserved from yeast cells to neurons posed a problem for neuroscientists. SNARE-mediated membrane fusion and protein secretion are rather slow, whereas the secretion of neurotransmitter molecules from synaptic vesicles in neurons occurs with millisecond precision and is tightly controlled by intracellular calcium ions, which can boost the vesicle fusion rate by up to 1 million times. Clearly, neuronal synapses could not rely only on SNAREs. They must contain a specialized protein machinery that boosts the somewhat sluggish SNARE machinery, thereby equipping synapses for their exquisite precision and speed.

When Rothman discovered the function of SNAREs in the early 1990s, Tom Südhof was well on his way to a systematic molecular cartography of synaptic communication between neurons and a functional analysis of synaptic-vesicle proteins — an endeavour that was co-pioneered with his long-term ally

Reinhard Jahn. Südhof identified and characterized numerous key components of the synaptic-vesicle fusion apparatus and of the parallel regulatory machinery that makes synaptic secretion so fast⁵. Of greatest importance, he identified synaptotagmin, and showed that this protein is the enigmatic calcium sensor that 'turbocharges' synaptic-vesicle secretion⁶.

But the story of this year's Nobel laureates does not end here. From single-cell organisms to humans, each and every cellular process depends on the cellular logistics of membrane trafficking and the secretion of cellular cargo. Not surprisingly then, diseases as diverse as tetanus, botulism, haemophagocytic lymphohistiocytosis, epilepsy and even schizophrenia have been shown — or are at least thought — to be caused by defects in proteins that control cellular trafficking. And this is just the tip of the iceberg. Interfering with these processes for therapeutic purposes seems only steps away. ■

Susan Ferro-Novick is at the Howard Hughes Medical Institute, Department of Cellular & Molecular Medicine, University of California, San Diego, La Jolla, California 92093-0668, USA.

Nils Brose is in the Department of Molecular Neurobiology, Max Planck Institute for Experimental Medicine, 37075 Göttingen, Germany.

e-mails: sfnovick@ucsd.edu; brose@em.mpg.de

1. Novick, P., Field, C. & Schekman, R. *Cell* **21**, 205–215 (1980).
2. Novick, P. & Schekman, R. *Proc. Natl Acad. Sci. USA* **76**, 1858–1862 (1979).
3. Balch, W. E., Dunphy, W. G., Braell, W. A. & Rothman, J. E. *Cell* **39**, 405–416 (1984).
4. Söllner, T. et al. *Nature* **362**, 318–324 (1993).
5. McMahon, H. T., Missler, M., Li, C. & Südhof, T. C. *Cell* **83**, 111–119 (1995).
6. Fernández-Chacón, R. et al. *Nature* **410**, 41–49 (2001).

PHYSICS

Endowing particles with mass

François Englert and Peter W. Higgs were awarded the Nobel Prize in Physics for the theoretical discovery of a mechanism that bestows mass on fundamental particles (see figure).

THE TRIUMPH OF A THEORY

by Ben Allanach

The proposal of the mass-giving mechanism was a coup for theoretical physics, and will remain a landmark for centuries to come. The standard model of particle physics successfully predicts a panoply of experimental data, some of them extremely precise, in very different contexts. Without the ideas of Robert

Brout, Englert¹, Higgs² and a few others, there is a fatal flaw in the standard model: it predicts that particles are massless, in clear contradiction to measurements. The mechanism that the researchers invented was the missing piece in a jigsaw puzzle, and the experimental detection of the Higgs boson at the Large Hadron Collider (LHC) at CERN, near Geneva, Switzerland, demonstrated that their ideas were correct.

As is often the case with scientific discoveries, this completed puzzle forms just one piece of a still larger one concerning quantum corrections to the Higgs-boson mass that occur at high energy scales. All particles receive quantum corrections to their masses from a boiling sea of other particles that pop in and out of existence, but normally the corrections are small and unproblematic. But for the Higgs boson, we have several corrections that are billions upon billions of times heavier than its measured mass (about 126 times the mass of a proton). So either quantum theory, which works so well in other contexts, is wrong, or we are missing a jigsaw piece.

There are exciting ideas for the missing piece that could solve this problem. One prominent idea, called supersymmetry, mathematically

cancels the huge corrections, and predicts a host of new particles to be discovered. We therefore await the restart of experiments at the LHC in 2015 with bated breath.

EXPERIMENTAL ENDORSEMENT

by Jonathan Butterworth

The first meeting to discuss what would become the LHC took place in 1984 in Lausanne, Switzerland. The 27-kilometre tunnel that now houses it was built for a previous accelerator, the Large Electron–Positron collider, which ran from 1989 to 2000. But as a result of even earlier discussions, the tunnel was built as large as possible to allow for future options³. These included later installation of a hadron collider, which became the LHC. In the 1990s, when researchers conceived the LHC's ATLAS and CMS particle detectors, with one of their major goals being to search for the Higgs boson, the required technology did not exist. A lengthy research and development programme was instigated to make sure that, when the time came, they could be built. Thousands of people have worked on them. These are just examples of the kind of long-term vision and investment needed to — using particle physicist David Miller's analogy⁴ — set off a rumour powerful enough to be heard.

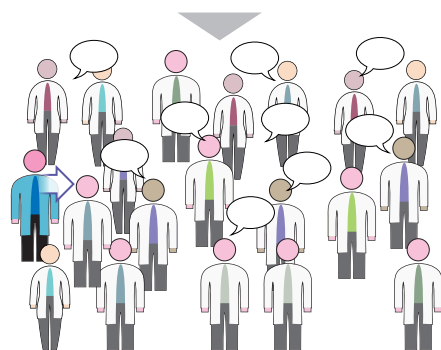
In the end, the result of this endeavour is simple^{5,6}. In a subset of recorded collisions between protons, there is a bump in the mass spectrum of pairs of photons and in the mass spectrum of four leptons (electrons and/or muons). That is the sign that we have managed to hit the background energy field of the Universe hard enough to make a wave in it. That wave is the Higgs boson. Many beautiful theoretical ideas have been proposed but consigned to oblivion, because they don't correspond to how the Universe works. Not so the mass mechanism proposed by Brout, Englert, Higgs and others — the boson is there! ■

Ben Allanach is in the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK. Jonathan Butterworth is in the Department of Physics and Astronomy, University College London, London WC1E 6BT, UK.
e-mails: b.c.allanach@damtp.cam.ac.uk; j.butterworth@ucl.ac.uk

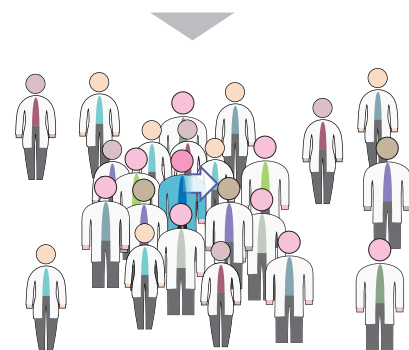
- Englert, F. & Brout, R. *Phys. Rev. Lett.* **13**, 321–323 (1964).
- Higgs, P. W. *Phys. Rev. Lett.* **13**, 508–509 (1964).
- LEP design report. CERN-LEP-84-01 (CERN, 1984).
- www.hep.ucl.ac.uk/~djm/higgsa.html
- The ATLAS Collaboration. *Phys. Lett. B* **716**, 1–29 (2012).
- The CMS Collaboration. *Phys. Lett. B* **716**, 30–61 (2012).

The mechanism proposed by Englert, Brout (deceased) and Higgs, among others, to explain the masses of fundamental particles involves the Higgs energy field and its associated particle, the Higgs boson. A rough analogy of the mechanism, based on an explanation by David J. Miller, goes like this:

1. A well-known scientist walks into a room full of uniformly spread physicists chatting with their fellows. The room is like space filled with the Higgs field.



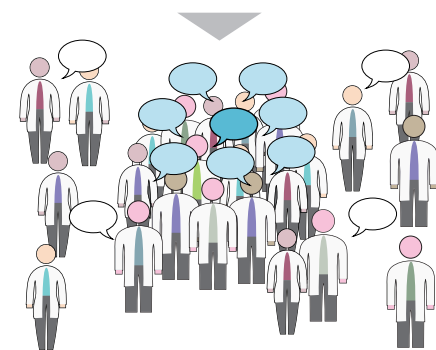
2. As he moves across the room, he attracts a 'mass' of admirers, much like a particle acquires mass by moving through the Higgs field.



3. Now imagine a rumour, about the well-known scientist, passing through the room.



4. The physicists will gather together to hear the rumour, creating a moving disturbance that is analogous to the Higgs boson.



Activation and allosteric modulation of a muscarinic acetylcholine receptor

Andrew C. Kruse^{1*}, Aaron M. Ring^{1,2*}, Aashish Manglik¹, Jianxin Hu³, Kelly Hu³, Katrin Eitel⁴, Harald Hübner⁴, Els Pardon^{5,6}, Celine Valant⁷, Patrick M. Sexton⁷, Arthur Christopoulos⁷, Christian C. Felder⁸, Peter Gmeiner⁴, Jan Steyaert^{5,6}, William I. Weis^{1,2}, K. Christopher Garcia^{1,2}, Jürgen Wess³ & Brian K. Kobilka¹

Despite recent advances in crystallography and the availability of G-protein-coupled receptor (GPCR) structures, little is known about the mechanism of their activation process, as only the β_2 adrenergic receptor (β_2 AR) and rhodopsin have been crystallized in fully active conformations. Here we report the structure of an agonist-bound, active state of the human M2 muscarinic acetylcholine receptor stabilized by a G-protein mimetic camelid antibody fragment isolated by conformational selection using yeast surface display. In addition to the expected changes in the intracellular surface, the structure reveals larger conformational changes in the extracellular region and orthosteric binding site than observed in the active states of the β_2 AR and rhodopsin. We also report the structure of the M2 receptor simultaneously bound to the orthosteric agonist iperoxo and the positive allosteric modulator LY2119620. This structure reveals that LY2119620 recognizes a largely pre-formed binding site in the extracellular vestibule of the iperoxo-bound receptor, inducing a slight contraction of this outer binding pocket. These structures offer important insights into the activation mechanism and allosteric modulation of muscarinic receptors.

Muscarinic acetylcholine receptors (M1–M5) are GPCRs that regulate the activity of a diverse array of central and peripheral functions in the human body, including the parasympathetic actions of acetylcholine¹. The M2 muscarinic receptor subtype has a key role in modulating cardiac function and many important central processes, such as cognition and pain perception¹. As it was among the first GPCRs to be purified² and cloned³, the M2 receptor has long served as a model system in GPCR biology and pharmacology. Muscarinic receptors have attracted particular interest owing to their ability to bind small-molecule allosteric modulators⁴. Because allosteric sites are often less conserved than the orthosteric binding site, some ligands binding to allosteric sites show substantial subtype selectivity^{5,6}. Such agents hold promise for the development of drugs for the treatment of conditions such as diseases of the central nervous system and for metabolic disorders. Although crystal structures were recently obtained for inactive states of the M2 and M3 muscarinic receptors^{7,8}, there are no structures of a GPCR bound to a drug-like allosteric modulator.

The binding of an agonist to the extracellular side of a GPCR results in conformational changes that enable the receptor to activate heterotrimeric G proteins. Despite the importance of this process, only the β_2 AR and rhodopsin have been crystallized in fully active conformations^{9–13}. Crystallization of active-state GPCRs has been challenging due to their inherent conformational flexibility and biochemical instability¹⁴. To understand the mechanistic details underlying GPCR activation and allosteric modulation better, we solved X-ray crystal structures of the M2 receptor bound to the high-affinity agonist iperoxo¹⁵ alone and in combination with LY2119620, a positive allosteric modulator.

Conformational selection of nanobodies

Initial crystallization attempts with M2 receptor bound to agonists were unsuccessful, probably due to the flexibility of the intracellular

receptor surface in the absence of a stabilizing protein. We thus sought to obtain a ‘G-protein mimetic’ nanobody for the M2 receptor, analogous to that used to facilitate crystallization of the β_2 AR in an active conformation¹¹. Llamas were immunized with M2 receptor bound to the agonist iperoxo, and a post-immune single variable domain (V_{HH}) nanobody complementary DNA library was constructed and displayed on the surface of yeast (Fig. 1a).

An essential component for the selection of active-state stabilizing nanobodies was simultaneous staining of yeast with both agonist and inverse-agonist occupied M2 receptor populations, which were distinguishably labelled with separate fluorophores. This allowed the use of fluorescence-activated cell sorting (FACS) to select those clones binding only agonist-occupied receptor (Fig. 1b; see Methods). To ensure that the different fluorophore-conjugated receptors represent distinct receptor populations requires that at least one receptor population must be bound to an exceptionally high-affinity or irreversible ligand. We therefore developed a covalent muscarinic receptor agonist for use in selection experiments. This has precedent in an acetylcholine mustard¹⁶, which is thought to react with the binding-site residue Asp 103^{3,32} (superscript numerals refer to the Ballesteros–Weinstein numbering system) to form a covalent adduct¹⁷. Accordingly, we synthesized an analogous ‘iperoxo mustard’, which we call FAUC123 (Supplementary Methods). We found that FAUC123 bound covalently and was able to induce activation of the M2 receptor (Extended Data Fig. 1), thereby allowing simultaneous staining of yeast with agonist- and antagonist-bound M2 receptor labelled with distinct fluorophores for each population.

After nine rounds of conformational selection, almost all remaining yeast cell clones preferentially bound FAUC123-occupied receptor (Fig. 1d). Three clones in particular, Nb9-1, Nb9-8 and Nb9-20

¹Department of Molecular and Cellular Physiology, Stanford University School of Medicine, 279 Campus Drive, Stanford, California 94305, USA. ²Department of Structural Biology, Stanford University School of Medicine, 299 Campus Drive, Stanford, California 94305, USA. ³Molecular Signaling Section, Laboratory of Bioorganic Chemistry, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland 20892, USA. ⁴Department of Chemistry and Pharmacy, Friedrich Alexander University, Schuhstrasse 19, 91052 Erlangen, Germany. ⁵Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium. ⁶Structural Biology Research Centre, VIB, Pleinlaan 2, B-1050 Brussels, Belgium. ⁷Drug Discovery Biology, Monash Institute of Pharmaceutical Sciences, and Department of Pharmacology, Monash University, Parkville, Victoria 3052, Australia. ⁸Neuroscience, Eli Lilly & Co., Indianapolis, Indiana 46285, USA.

*These authors contributed equally to this work.

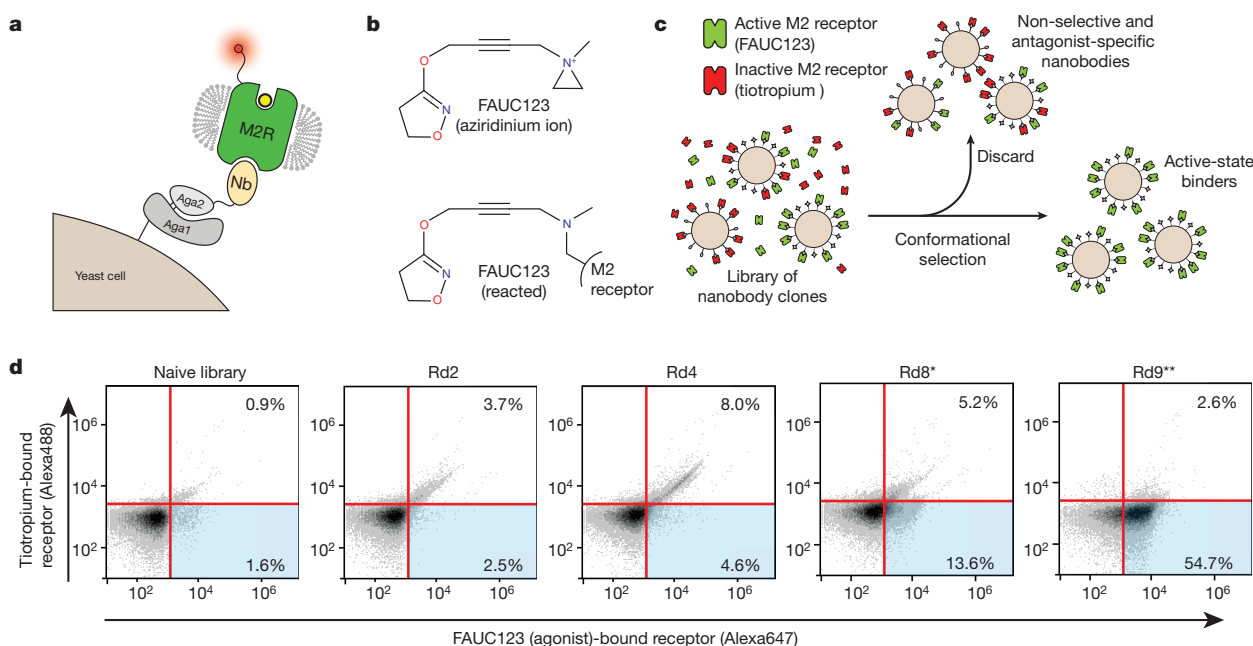


Figure 1 | Isolation of Nb9-8. **a**, Nanobodies from a llama immunized with M2 receptor were displayed on yeast as an amino-terminal fusion to Aga2, and subjected to magnetic selection to enrich clones that bind preferentially to agonist-occupied receptor. **b**, For selections, an aziridinium ion derivative of iperoxo called FAUC123 was synthesized, allowing covalent modification of the receptor. **c**, Yeast were stained simultaneously with agonist-occupied M2 receptor and antagonist-occupied receptor labelled with distinct fluorophores.

d, Yeast from each selection round (Rd1–9) were stained in this manner to assess selection progress, showing a clear enrichment first for non-selective binders (upper-right quadrants) followed by specific enrichment for agonist-preferring clones (lower-right quadrants). A single asterisk indicates a selection round using conformational selection. Two asterisks indicate a selection round using FACS.

(Fig. 2a; see Methods), showed strong, conformationally selective staining on yeast (Fig. 2b). All three nanobodies enhanced agonist affinity (Fig. 2c), indicating that they stabilize active states of the receptor. Nb9-8 was the most potent, with a half-maximum effective concentration (EC_{50}) of approximately 100 nM. At high concentrations, Nb9-8 enhanced the affinity of the M2 receptor for iperoxo to almost the same extent as that observed in the presence of the heterotrimeric G protein G_i (Fig. 2d).

M2 receptor was purified in the presence of 10 μ M iperoxo, and we obtained crystals of iperoxo-bound M2 receptor in complex with Nb9-8 by lipidic mesophase crystallography. The structure was solved by microdiffraction at Advanced Photon Source beamline 23ID-D (Extended Data Table 1). Supplementing the optimized crystallization conditions with the positive allosteric modulator LY2119620 yielded crystals of M2 receptor simultaneously bound to both iperoxo and the modulator (see Methods). For all crystallization work, the agonist iperoxo was used rather than acetylcholine, as the latter is of lower affinity and is prone to hydrolysis.

Cytoplasmic changes on activation

A key feature of GPCR activation is an outward movement of the intracellular portion of transmembrane (TM) helices 5 and 6, creating a cavity large enough to accommodate the carboxy terminus of the G protein α -subunit^{10,13}. Although several GPCRs have been crystallized in complex with agonists, only the β_2 AR and rhodopsin show a fully active state with adequate space to allow G-protein binding (Extended Data Fig. 2). As anticipated on the basis of functional studies (Fig. 2), Nb9-8 binds to the intracellular surface of the receptor (Fig. 3a). There is a significant outward displacement at the intracellular side of TM6, together with a smaller outward movement of TM5 and a rearrangement of TM7 around the NPXXY motif (Fig. 3b, d).

Like the active states of rhodopsin and the β_2 AR, the active M2 receptor shows rearrangements of the highly conserved DRY motif at the intracellular side of TM3 and the NPXXY motif in TM7 (Fig. 3c, d).

In the active state of M2, Arg 121^{3,50} of the DRY motif adopts an extended conformation virtually identical to that seen in metarhodopsin II and the β_2 AR- G_s complex (Fig. 3c, e), and Asp 120^{3,49} is stabilized by a hydrogen bond with Asn 58^{2,39} (Fig. 3c). To assess the importance of Asn 58^{2,39} for stabilization of the active conformation, we mutated it to alanine. The resulting mutant displayed normal ligand-binding properties, but impaired ability to activate G protein (Extended Data Fig. 3a and Extended Data Table 2). Hence, it is likely that Asn 58^{2,39} either directly stabilizes the active conformation, or engages in direct interactions with G protein.

Similar to the DRY motif, the NPXXY region in TM7 shows significant rearrangements on activation (Fig. 3d). Most striking is a partial ‘unwinding’ of TM7 around Tyr 440^{7,53}. This positions Tyr 440^{7,53} of the NPXXY motif in close proximity to the highly conserved residue Tyr 206^{5,58} (Fig. 3d). Although these two residues are not close enough to interact directly, their proximity may allow formation of a water-mediated hydrogen bond, as seen in the active-state structures of the β_2 AR¹⁸ and rhodopsin¹². Indeed, the position of these two tyrosine residues is highly similar in the active structures of rhodopsin, β_2 AR and the M2 receptor (Fig. 3f), indicating that this feature represents a hallmark of GPCR activation. In addition, a molecular dynamics study recently predicted that Tyr 206^{5,58} and Tyr 440^{7,53} interact in the active conformation of the M2 receptor¹⁹, although this model was in other ways dissimilar from the structures presented here.

To assess the importance of this interaction for M2 receptor activation, we mutated Tyr 206^{5,58} to phenylalanine, eliminating its ability to interact with Tyr 440^{7,53} via a bridging water molecule. The Y206F mutant receptor could no longer be activated by acetylcholine (Extended Data Fig. 3a) and gave only a very weak functional response on treatment with iperoxo. In addition, agonist affinity was reduced by greater than tenfold (Extended Data Table 2), whereas antagonist binding was largely unaffected. These results indicate that the Tyr 206^{5,58}–Tyr 440^{7,53} interaction stabilizes the active conformation of the receptor in a manner reminiscent of the ‘ionic lock’ interaction²⁰, which stabilizes the inactive conformation of family A GPCRs.

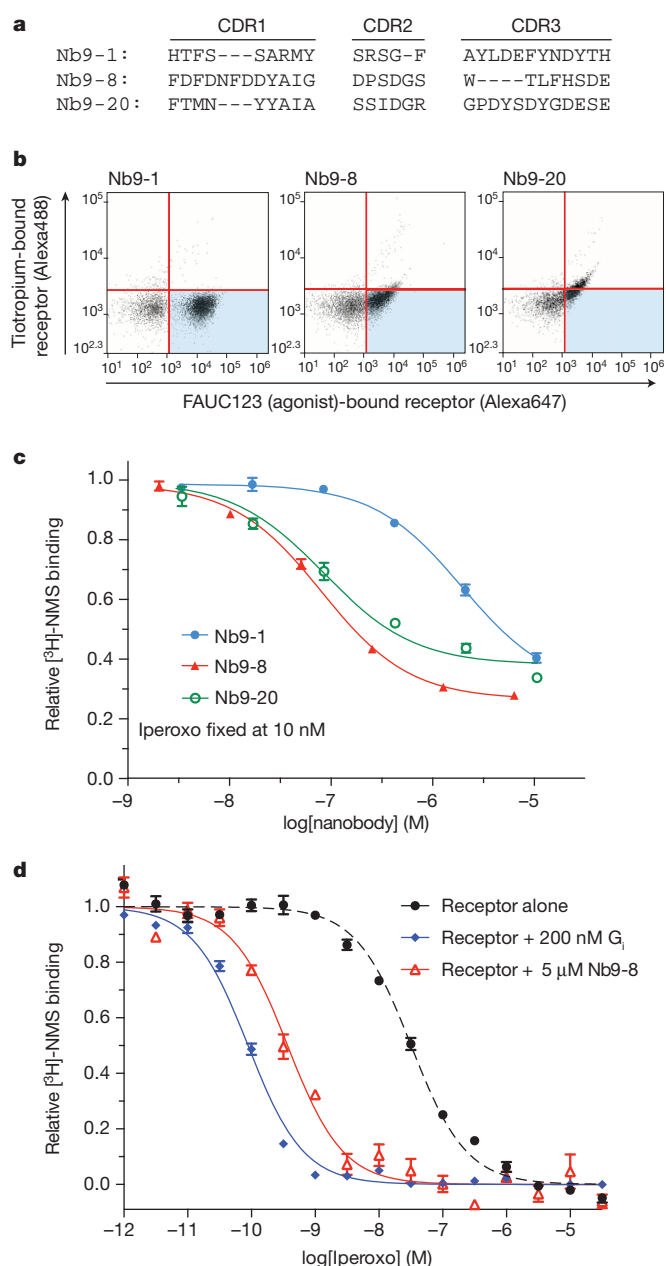


Figure 2 | M2 active-state-specific nanobodies. **a**, Three nanobodies were selected for detailed characterization, each with entirely unique complementarity determining region (CDR) sequences. These three nanobodies were expressed on the surface of yeast, and characterized by flow cytometry staining with FAUC123-bound (that is, agonist bound) M2 receptor and tiotropium-bound (that is, antagonist occupied) receptor. **b**, Each of the three clones displayed a preference for agonist-occupied receptor to varying degrees. **c**, Purified nanobodies were tested in a concentration–response assay for their ability to suppress [3 H]-NMS binding to the M2 receptor in the presence of 10 nM (IC_{50}) iperoxo, with Nb9-8 being the most potent clone. **d**, Like the G protein G_i , Nb9-8 caused a substantial enhancement of iperoxo affinity in a competition binding assay. Panels **c** and **d** are representative of at least three experiments performed in triplicate, and the data and error bars represent the mean \pm s.e.m.

Activation mechanism

Whereas activation of β_2 AR and rhodopsin is associated with modest conformational changes in the orthosteric ligand-binding site, marked structural changes are observed in the M2 receptor. The activated M2 receptor shows a small orthosteric binding site, which completely occludes the agonist iperoxo from solvent (Fig. 4a, b). Indeed, the muscarinic inverse agonist quinuclidinyl benzilate (QNB) is too large

to be accommodated in this binding cavity, perhaps accounting for its ability to suppress basal activity of the M2 receptor.

Within the active orthosteric binding pocket, the agonist iperoxo adopts a bent conformation (Fig. 4c and Extended Data Fig. 4). Trans-membrane helices 5, 6 and 7 move inward, towards the agonist, in the active M2 conformation. TM3, in contrast, undergoes a slight rotation about its axis, but has almost no inward motion towards the ligand. The largest differences between inactive and active states of the M2 receptor involve TM6, where an inward movement of 2 Å at the α -carbon of Asn 404^{6,52} allows for formation of a hydrogen bond between its side chain and iperoxo.

Despite these activation-related structural changes, polar contacts between the agonist iperoxo and the receptor resemble those with QNB bound to the inactive M2 receptor. In particular, the conserved Asp 103^{3,32} serves as a counter-ion to the ligand amine in both cases, and Asn 404^{6,52} engages in hydrogen bonding with both ligands. The smaller size of iperoxo relative to QNB results in more limited hydrophobic contacts, however. This is particularly true along TMs5, which engages the phenyl rings of QNB, but makes more limited hydrophobic contact with iperoxo in the active receptor conformation.

The hydrogen bond between Asn 404^{6,52} and the iperoxo isoxazoline oxygen is analogous to the hydrogen bond between this residue and the QNB carbonyl in the inactive receptor state; however, the smaller size of iperoxo necessitates an inward motion of TM6 (Fig. 4d, e). To investigate the role of this hydrogen bond in receptor activation, we mutated Asn 404^{6,52} to glutamine, which, due to the longer side chain, would allow TM6 to form a hydrogen bond with iperoxo in the inactive receptor. Consistent with a previous mutagenesis study²¹, the N404Q mutant receptor failed to bind detectable amounts of [3 H]-NMS, but retained the ability to bind [3 H]-QNB specifically, although with 163-fold reduced affinity (Extended Data Table 2). Similarly, the binding affinities for acetylcholine and iperoxo were reduced, and although the N404Q mutant was able to activate G protein in response to both iperoxo and acetylcholine, the concentration–response curves were shifted to the right by about 100-fold (Extended Data Fig. 3a), probably due to the reduced agonist-binding affinities. Nevertheless, it remains possible that a structural reorientation of Asn 404^{6,52} also contributes to M2 receptor activation.

Like Asn 404^{6,52}, Asp 103^{3,32} has a central role in receptor binding to iperoxo, engaging the trimethyl ammonium ion. Cation- π interactions with Tyr 104^{3,33}, Tyr 403^{6,51} and Tyr 426^{7,39} form an aromatic lid over the ligand amine (Fig. 4f). To assess the contribution of Asp 103^{3,32} to receptor activation, we generated and analysed the D103E mutant M2 receptor, which abolished agonist-induced M2 receptor activation (Extended Data Fig. 3a). The D103E mutant receptor bound [3 H]-NMS with wild-type-like affinity but showed greatly reduced affinities for acetylcholine (\sim 120-fold) and iperoxo (\sim 380-fold) (Extended Data Table 2), indicating that Asp 103^{3,32} recognition of the ligand cation has a critical role in both agonist binding and receptor activation.

In the active state of the M2 receptor, the inward motion of the upper portion of TM6 allows Tyr 403^{6,51} to form a hydrogen bond with Tyr 104^{3,33}, which in turn forms a hydrogen bond to Tyr 426^{7,39} (Fig. 4f), resulting in closure of the aforementioned tyrosine lid over the agonist. Hydrogen bonding of this lid seems to be an important feature of agonist binding and activation in muscarinic receptors: mutation of any of the three tyrosines to Phe leads to impaired agonist binding in the homologous M3 muscarinic receptor²², and mutation of Tyr 104^{3,33} and Tyr 403^{6,51} in the M2 receptor has a similar effect^{23,24}. It should be noted that the structure of active M2 receptor bound to other agonists, including acetylcholine, might show differences as compared to the iperoxo-bound structure presented here.

Allosteric modulation

Muscarinic receptors have long served as important model systems for understanding allosteric modulation of GPCR signalling^{5,6,25}. The structures of the inactive M2 and M3 receptors confirmed that these

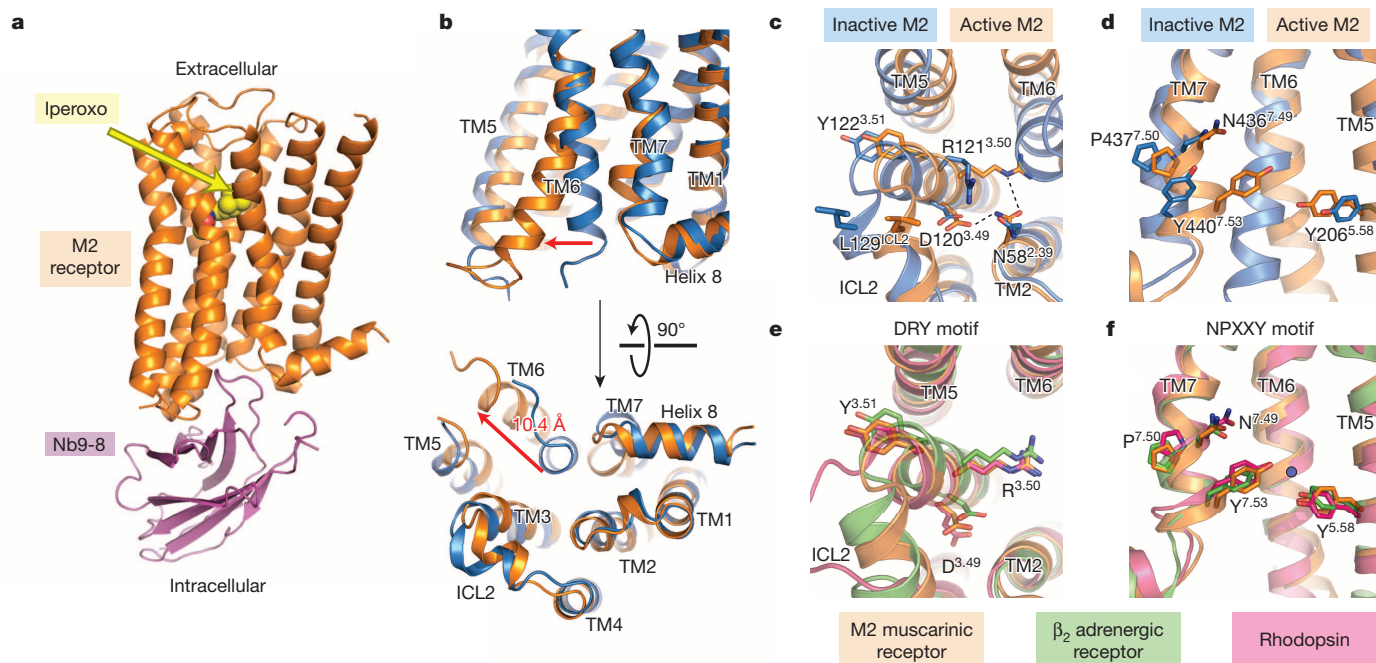


Figure 3 | Intracellular changes on activation of the M2 receptor. **a**, The overall structure of the active-state M2 receptor (orange) in complex with the orthosteric agonist iperoxo and the active-state stabilizing nanobody Nb9-8 is shown. **b**, Compared to the inactive structure of the M2 receptor (blue), transmembrane helix 6 (TM6) is substantially displaced outward, and TM7 has moved inward. Together, these motions lead to the formation of the

receptors possess a large extracellular vestibule, which has been shown to bind to allosteric modulators^{26,27}. Situated directly above (that is, extracellular to) the orthosteric site, this cavity also shows a substantial contraction upon activation of the M2 receptor due to the rotation of TM6 (Fig. 4b). The motion of TM6 thus provides a structural link among three regions of the receptor: the extracellular vestibule, the orthosteric binding pocket, and the intracellular surface. The structural coupling of these three regions probably accounts for the fact that allosteric modulators can affect the affinity and efficacy of orthosteric ligands and can also directly activate G proteins as allosteric agonists²⁸.

For a better understanding of how allosteric modulators act at GPCRs, we crystallized the iperoxo-occupied M2 receptor with LY2119620, a positive allosteric modulator (Fig. 5a). This agent has not been studied previously, so we characterized its affinity for the M2 receptor and its allosteric interaction with iperoxo (see Supplementary Methods). Radioligand binding assays revealed that LY2119620 has similar pharmacological properties to its congener, LY2033298 (ref. 29) (Extended Data Fig. 3b and Extended Data Table 3). It shows strong positive cooperativity with iperoxo, and mild negative cooperativity with the inverse agonist [³H]-NMS. Whereas LY2119620 enhances the affinity of the M2 receptor for iperoxo, it does not significantly change the efficacy of this orthosteric agonist (Extended Data Table 3). We also observed that LY2119620 is capable of directly activating the M2 receptor, albeit with low potency and efficacy relative to iperoxo (Extended Data Table 3).

Crystals of the M2 receptor bound to LY2119620 grew under identical conditions to those without the modulator, and the structure revealed unambiguous electron density for LY2119620 in the extracellular vestibule (Extended Data Fig. 5). The modulator is positioned directly above the orthosteric agonist (Fig. 5b), and it engages in extensive interactions with the extracellular vestibule. Specifically, the aromatic rings of the modulator are situated directly between Tyr 177^{ECL2} (where ECL2 indicates extracellular loop 2) and Trp 422^{7.35}, forming a three-layered aromatic stack. Importantly, a previous mutagenesis

G-protein-binding site. **c, d**, Conserved motifs likewise show substantial changes on activation, and adopt conformations similar to those seen in the two other active GPCR structures (**e, f**). In particular, an interaction between two conserved tyrosines (Tyr^{5.58} and Tyr^{7.53}) is probably mediated by a water molecule (blue circle), as seen in the high-resolution structure of the active β_2 AR (ref. 18).

study implicated Tyr 177^{ECL2} as a likely contact for the LY2119620 congener, LY2033298, at the M2 muscarinic receptor²⁹. Several polar interactions are also seen (Fig. 5c). In particular, Tyr 80^{2.61}, Asn 410^{6.58} and Asn 419^{ECL3} form hydrogen bonds to the modulator, and Glu 172^{ECL2} engages in a charge–charge interaction with the ligand piperidine. LY2119620 binds at a site directly superficial to the orthosteric site, separated only by the tyrosine lid, with Tyr 426^{7.39} interacting with both ligands.

The structure of the M2–iperoxo–LY2119620 complex is largely the same as that of receptor and agonist without LY2119620, indicating that the allosteric binding site is largely pre-formed in the presence of agonist. The extracellular vestibule shows a slight additional contraction around the allosteric ligand (Extended Data Fig. 6). This subtle change stands in contrast to the substantial closure of the extracellular vestibule in the two active structures relative to the inactive conformation (Fig. 5d). A notable exception is Trp 422^{7.35}, which adopts a vertical conformation in the presence of LY2119620 and a horizontal conformation with iperoxo alone (Extended Data Fig. 6b). The vertical conformation of this residue in the M2–iperoxo–LY2119620 complex allows it to engage in an aromatic stacking interaction with the modulator, consistent with mutagenesis results implicating Trp 422^{7.35} in the binding of other allosteric modulators³⁰. The effect of mutagenesis of Trp 422^{7.35} on LY2119620 affinity has not been tested, however. Closure of the LY2119620 binding site in the agonist-bound M2 receptor allows far more extensive interactions with the modulator than the inverse agonist-bound conformation (Fig. 5e), probably accounting for the ability of the modulator to enhance agonist binding affinity by preferentially slowing agonist dissociation.

The closed, active conformation of the extracellular vestibule is largely the consequence of the inward motion of TM6, which directly contacts the allosteric modulator, the orthosteric agonist, and probably the G protein as well. Stabilization of the closed extracellular vestibule by LY2119620 and other allosteric modulators may directly stabilize the open, active conformation of the intracellular side of TM6, accounting for the phenomenon of allosteric agonism in addition to positive

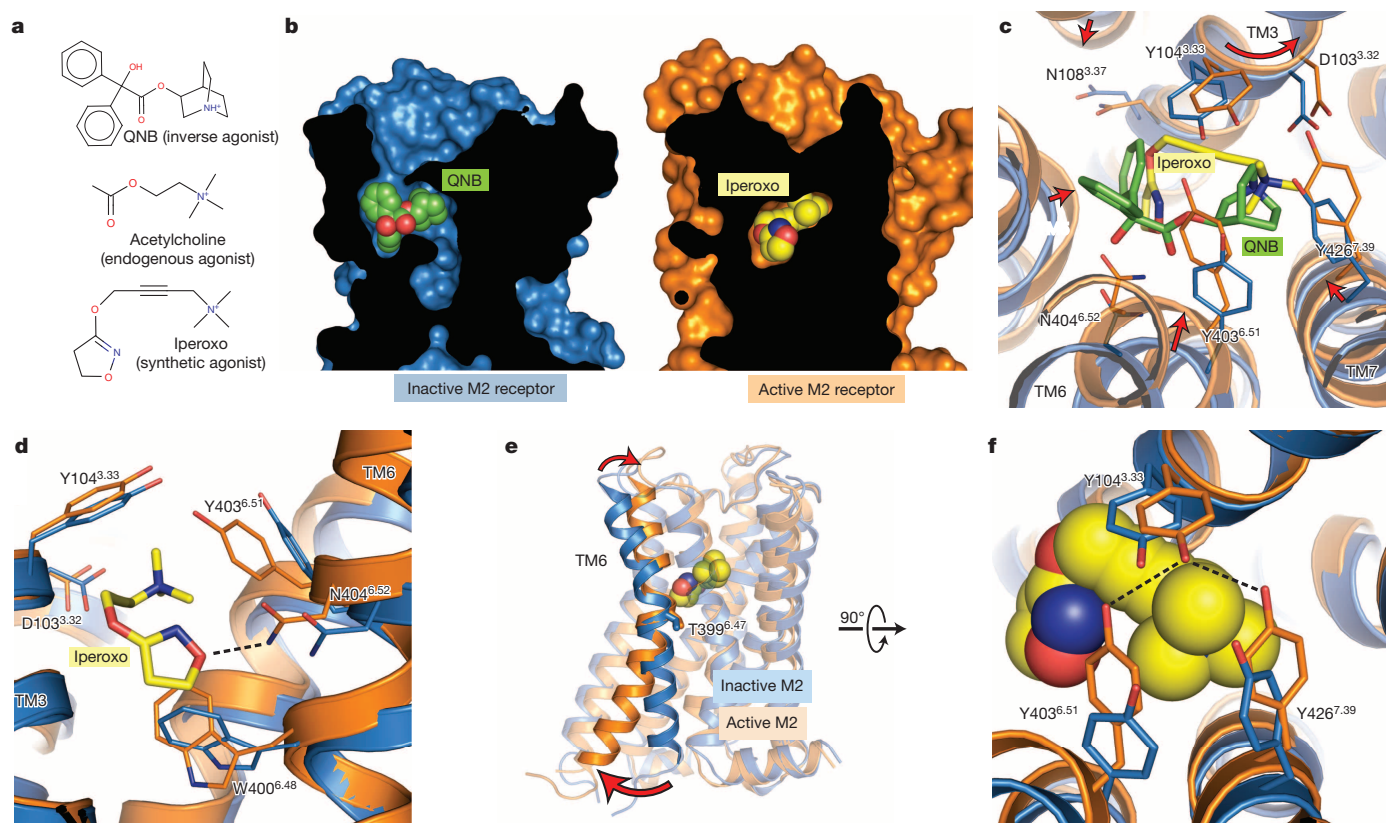


Figure 4 | Orthosteric ligand-binding site. **a**, Orthosteric ligands used for crystallization of inactive and active M2 receptor are shown. **b**, Cross-sections through the receptor are shown, with the interior in black. In the inactive conformation, the receptor (blue, at left) partially encloses the antagonist QNB, while the active conformation receptor encloses the agonist entirely, such that it is completely buried within the receptor (orange, at right). **c**, Conformational changes within the ligand-binding pocket are shown from the extracellular side,

with changes highlighted as red arrows. **d**, A side view shows the inward motion of TM6, which is required for the formation of a hydrogen bond between Asn 404^{6.52} and the agonist iperoxo. **e**, Activation thus involves a pivot of TM6, which moves inward in the orthosteric site and outward at the intracellular side. **f**, The closure of the binding pocket allows the formation of a hydrogen-bonded tyrosine lid, located superficial to the agonist.

cooperativity with orthosteric agonists. However, although the differences in TM6 between inactive and active structures can be described as a rigid-body motion, we cannot exclude the possibility that TM6 is flexible, allowing independent conformational changes in the G-protein binding site, the orthosteric site and the extracellular vestibule.

Conclusions

The structures presented here offer insights into the structural basis for muscarinic receptor activation and allosteric modulation by a drug-like molecule. In contrast to rhodopsin and the β_2 AR, extensive changes are seen in the orthosteric binding site and in the extracellular

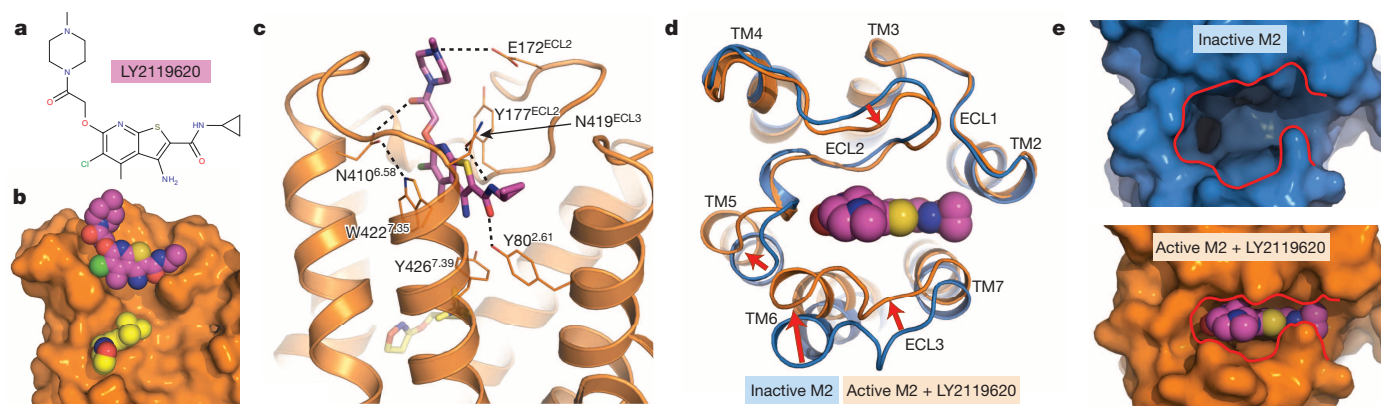


Figure 5 | Structure of a GPCR allosteric modulator complex. **a**, The M2 receptor occupied by the orthosteric agonist iperoxo was crystallized in complex with the positive allosteric modulator LY2119620. **b**, The allosteric ligand binds to the extracellular vestibule just above the orthosteric agonist. A cross-section through the membrane plane shows the relative positions of the two ligands. **c**, Several polar contacts are involved in LY2119620 binding, in

addition to extensive aromatic stacking interactions with Trp 422^{7.35} and Tyr 177^{ECL2}. **d**, Upon activation, the M2 receptor undergoes substantial conformational changes in the extracellular surface, leading to a contraction of the extracellular vestibule. **e**, This creates a binding site that fits tightly around the allosteric modulator, which would otherwise be unable to interact extensively with the extracellular vestibule in the inactive receptor conformation.

vestibule upon M2 receptor activation. The structure of active M2 receptor bound to the allosteric modulator LY2119620 definitively establishes the extracellular vestibule as an allosteric binding site, and shows that the allosteric modulator induces few additional structural changes as compared to those seen with orthosteric agonist alone. The structures presented here offer only a single view of an active muscarinic receptor; more work will be required to identify additional active states that may exist. Nonetheless, the information presented here provides a structural framework for future studies of GPCR activation and allostery, and may facilitate the development of novel therapeutics.

METHODS SUMMARY

The human M2 muscarinic receptor was expressed in Sf9 insect cells and purified to homogeneity by nickel affinity chromatography, followed by Flag affinity and size exclusion chromatography. The nanobody Nb9-8 was identified by yeast surface display using a library derived from peripheral blood lymphocytes of a llama immunized with purified, iperoxo-occupied M2 receptor. Recombinant Nb9-8 was expressed in the periplasm of *Escherichia coli* strain BL21(DE3), and purified by nickel affinity chromatography followed by size exclusion chromatography. Crystallography was performed using lipidic mesophase methods, and data were collected by X-ray microdiffraction at Advanced Photon Source GM/CA beamlines 23ID-B and 23ID-D.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 July; accepted 3 October 2013.

Published online 20 November 2013.

- Wess, J., Eglen, R. M. & Gautam, D. Muscarinic acetylcholine receptors: mutant mice provide new insights for drug development. *Nature Rev. Drug Discov.* **6**, 721–733 (2007).
- Peterson, G. L., Herron, G. S., Yamaki, M., Fullerton, D. S. & Schimerlik, M. I. Purification of the muscarinic acetylcholine receptor from porcine atria. *Proc. Natl Acad. Sci. USA* **81**, 4993–4997 (1984).
- Kubo, T. *et al.* Primary structure of porcine cardiac muscarinic acetylcholine receptor deduced from the cDNA sequence. *FEBS Lett.* **209**, 367–372 (1986).
- Mohr, K., Trankle, C. & Holzgrabe, U. Structure/activity relationships of M2 muscarinic allosteric modulators. *Receptors Channels* **9**, 229–240 (2003).
- Digby, G. J., Shirey, J. K. & Conn, P. J. Allosteric activators of muscarinic receptors as novel approaches for treatment of CNS disorders. *Mol. Biosyst.* **6**, 1345–1354 (2010).
- Keov, P., Sexton, P. M. & Christopoulos, A. Allosteric modulation of G protein-coupled receptors: a pharmacological perspective. *Neuropharmacology* **60**, 24–35 (2011).
- Haga, K. *et al.* Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482**, 547–551 (2012).
- Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482**, 552–556 (2012).
- Choe, H. W. *et al.* Crystal structure of metarhodopsin II. *Nature* **471**, 651–655 (2011).
- Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011).
- Rasmussen, S. G. *et al.* Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469**, 175–180 (2011).
- Deupi, X. *et al.* Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II. *Proc. Natl Acad. Sci. USA* **109**, 119–124 (2012).
- Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455**, 497–502 (2008).
- Nygaard, R. *et al.* The dynamic process of β_2 -adrenergic receptor activation. *Cell* **152**, 532–542 (2013).
- Kloekner, J., Schmitz, J. & Holzgrabe, U. Convergent, short synthesis of the muscarinic superagonist iperoxo. *Tetrahed. Lett.* **51**, 3470–3472 (2010).
- Hudgins, P. M. & Stubbins, J. F. A comparison of the action of acetylcholine and acetylcholine mustard (chloroethylmethylaminoethyl acetate) on muscarinic and nicotinic receptors. *J. Pharmacol. Exp. Ther.* **182**, 303–311 (1972).
- Spalding, T. A., Birdsall, N. J., Curtis, C. A. & Hulme, E. C. Acetylcholine mustard labels the binding site aspartate in muscarinic acetylcholine receptors. *J. Biol. Chem.* **269**, 4092–4097 (1994).
- Ring, A. M. *et al.* Adrenaline-activated structure of the β_2 -adrenoceptor stabilized by an engineered nanobody. *Nature* **502**, 575–579 (2013).
- Miao, Y., Nichols, S. E., Gasper, P. M., Metzger, V. T. & McCammon, J. A. Activation and dynamic network of the M2 muscarinic receptor. *Proc. Natl Acad. Sci. USA* **110**, 10982–10987 (2013).
- Ballesteros, J. A. *et al.* Activation of the β_2 -adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276**, 29171–29177 (2001).
- Heitz, F. *et al.* Site-directed mutagenesis of the putative human muscarinic M2 receptor binding site. *Eur. J. Pharmacol.* **380**, 183–195 (1999).
- Wess, J., Maggio, R., Palmer, J. R. & Vogel, Z. Role of conserved threonine and tyrosine residues in acetylcholine binding and muscarinic receptor activation. A study with m3 muscarinic receptor point mutants. *J. Biol. Chem.* **267**, 19313–19319 (1992).
- Vogel, W. K., Sheehan, D. M. & Schimerlik, M. I. Site-directed mutagenesis on the m2 muscarinic acetylcholine receptor: the significance of Tyr 403 in the binding of agonists and functional coupling. *Mol. Pharmacol.* **52**, 1087–1094 (1997).
- Gregory, K. J., Hall, N. E., Tobin, A. B., Sexton, P. M. & Christopoulos, A. Identification of orthosteric and allosteric site mutations in M2 muscarinic acetylcholine receptors that contribute to ligand-selective signaling bias. *J. Biol. Chem.* **285**, 7459–7474 (2010).
- De Amici, M., Dallanoc, C., Holzgrabe, U., Trankle, C. & Mohr, K. Allosteric ligands for G protein-coupled receptors: a novel strategy with attractive therapeutic opportunities. *Med. Res. Rev.* **30**, 463–549 (2010).
- Gregory, K. J., Sexton, P. M. & Christopoulos, A. Allosteric modulation of muscarinic acetylcholine receptors. *Curr. Neuropharmacol.* **5**, 157–167 (2007).
- Bock, A. *et al.* The allosteric vestibule of a seven transmembrane helical receptor controls G-protein coupling. *Nature Commun.* **3**, 1044 (2012).
- May, L. T. *et al.* Structure-function studies of allosteric agonism at M2 muscarinic acetylcholine receptors. *Mol. Pharmacol.* **72**, 463–476 (2007).
- Valant, C., Felder, C. C., Sexton, P. M. & Christopoulos, A. Probe dependence in the allosteric modulation of a G protein-coupled receptor: implications for detection and validation of allosteric ligand effects. *Mol. Pharmacol.* **81**, 41–52 (2012).
- Prilla, S., Schrobang, J., Ellis, J., Holtje, H. D. & Mohr, K. Allosteric interactions with muscarinic acetylcholine receptors: complex role of the conserved tryptophan M2422Trp in a critical cluster of amino acids for baseline affinity, subtype selectivity, and cooperativity. *Mol. Pharmacol.* **70**, 181–193 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge support from the National Science Foundation (graduate fellowship to A.C.K., and Award 1223785 to B.K.K.), the Stanford Medical Scientist Training Program (A.M. and A.M.R.), the American Heart Association (A.M.), the Ruth L. Kirschstein National Research Service Award (A.M.R.), National Institutes of Health grants NS02847123 and GM08311806 (B.K.K.), the Mathers Foundation (B.K.K., W.I.W. and K.C.G.), the Deutsche Forschungsgemeinschaft for the grant GM 13/10-1 (K.E., H.H., P.G.), the National Health and Medical Research Council (NHMRC) of Australia program grant 519461 (P.M.S. and A.C.), NHMRC Principal Research Fellowships (P.M.S. and A.C.), and the Howard Hughes Medical Institute (K.C.G.). This work was supported in part by the Intramural Research Program, NIDDK, NIH, US Department of Health and Human Services (J.H., K.H. and J.W.). We thank K. Leach for performing ERK assays, and B. Davie and P. Scammells for synthesis of iperoxo. We thank H. Xiao, C. H. Croy and D. A. Schober for functional characterization of LY2119620. We thank T. S. Kobilka for preparation of affinity chromatography reagents and F. S. Thian for help with cell culture.

Author Contributions A.C.K. expressed and purified M2 receptor for yeast display and crystallographic experiments, performed crystallization, data collection, and structure refinement, and performed radioligand binding assays to validate nanobody activity. A.C.K., A.M.R. and A.M. designed experiments to identify nanobodies by yeast display. A.M.R. performed all yeast selections, and expressed and purified Nb9-8 and other nanobodies. J.H. and K.H. performed site-directed mutagenesis and characterization of resulting mutants. K.E. synthesized FAUC123. H.H. performed cell assays and radioligand binding to characterize FAUC123. C.V. performed pharmacological characterization of LY2119620. P.M.S. and A.C. supervised pharmacological characterization of LY2119620. C.C.F. designed key solubility, physical chemistry and ligand analysis to select LY2119620 as an appropriate co-crystallization candidate for the M2 receptor. P.G. supervised synthesis and characterization of FAUC123. E.P. and J.S. performed llama immunization, cDNA production, and performed selections by phage display. W.I.W. supervised structure refinement. K.C.G. supervised yeast selection experiments. J.W. supervised mutagenesis experiments and analysed results. B.K.K. provided overall project supervision, and with A.C.K., A.M.R. and A.M. wrote the manuscript with assistance from A.C. and J.W.

Author Information Coordinates and structure factors for the active M2 receptor in complex with Nb9-8 and iperoxo are deposited in the Protein Data Bank under accession code 4MQS, and the coordinates and structure factors of the same complex bound additionally to the allosteric modulator LY2119620 are deposited under accession code 4MQT. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.K.K. (kobilka@stanford.edu).

METHODS

Determination of M2 activation via inositol phosphate assays. Agonist-induced activation of the human M2 muscarinic receptor was studied in inositol phosphate (IP) accumulation assays as described³¹. For M2 activation studies, COS-7 cells were transiently co-transfected with cDNAs encoding the human M2 receptor (Missouri S&T cDNA Resource Center) and the hybrid G protein $G_{\alpha_{q15}}$ (G_{α_q} protein with the last five amino acids at the C terminus replaced by the corresponding sequence of G_{α_i} ; gift from The J. David Gladstone Institutes)³². Twenty-four hours after transfection, cells were transferred into 24-well plates at a density of 100,000 cells per well in a volume of 270 μ l. After addition of 30 μ l of *myo*-[³H]inositol (specific activity = 22.5 Ci mmol⁻¹, PerkinElmer), cells were incubated for 15 h. Then, medium was aspirated, the cells were washed with serum-free medium supplemented with 10 mM LiCl, and test compounds (diluted in serum-free medium supplemented with 10 mM LiCl) were added at 37 °C for 60 min. Cells were then lysed by adding 150 μ l of ice-cold 0.1 M NaOH for 5 min. After neutralization with 50 μ l of 0.2 M formic acid, the cell extract was diluted in buffer (5 mM sodium tetraborate, 0.5 mM Na-EDTA) and separated by anion-exchange chromatography using an AG1-X8 resin (Bio-Rad). After washing with water and elution buffer A (5 mM sodium tetraborate, 60 mM sodium formate) and again with water, total IP was eluted with 2.5 ml elution buffer B (1.0 M ammonium formate) and directly collected into scintillation counting vials. Radioactivity was measured by scintillation counting after adding 2.5 ml of Emulsifier-Safe (PerkinElmer). Data were analysed by normalizing disintegrations per minute (d.p.m.) values with 0% for the non-stimulated receptor and 100% for the full effect of the reference iperexo. Concentration–response curves were fitted by nonlinear regression using the Graphpad Prism 5 software.

Irreversible activation of the M2 receptor was tested at 1 nM FAUC123 in comparison to the reversible ligand iperexo (1 nM). After incubation for 30 min, the antagonist atropine (1 μ M) was added to one-half of the sample (buffer was added to the other half) and incubations were continued for an additional 90 min. Total IP accumulation was determined as described above.

LY2119620 pharmacology. To characterize the allosteric interaction between LY2119620 and iperexo, we performed radioligand binding and cellular functional assays at the wild-type human M2 muscarinic receptor stably expressed in a CHO FlpIn cell line. Increasing concentrations of LY2119620 caused a modest reduction in the specific binding of the orthosteric antagonist, [³H]-NMS, indicating weak negative cooperativity, but robustly enhanced the potency of iperexo to compete for [³H]-NMS binding, indicating positive cooperativity with the agonist (Extended Data Fig. 3b). Application of an allosteric ternary complex model^{33,34} to these data yielded the values shown in Extended Data Table 3 for ligand affinity and cooperativities with agonist and antagonist. We then investigated the functional effect of LY2119620 on M2 muscarinic receptor signalling via monitoring receptor-mediated [³⁵S]-GTP γ S binding to activated G proteins, or phosphorylation of ERK1/2 (pERK/2). [³⁵S]-GTP γ S binding was chosen as a proximal measure of receptor activation, whereas the pERK1/2 assay was chosen because it measures a downstream response that is also a point of convergence of multiple cellular pathways, some of them potentially G-protein independent. In both instances, LY2119620 caused receptor activation in its own right, indicating that the modulator can act as an allosteric agonist, while simultaneously enhancing the potency of iperexo (Extended Data Fig. 3b). Application of an operational model of allostery³⁵ to these data yielded the parameter values shown in Extended Data Table 3. Comparison of the binding and functional data indicated that there was no significant difference between any of the pK_B estimates of the affinity of LY2119620 for the allosteric site on the free receptor between assays. There was also no significant difference between the cooperativity factors with iperexo across the assays, indicating that the molecular mechanism of action of LY2119620 is consistent with positive modulation of agonist affinity only, with minimal additional effects on agonist efficacy. This is in contrast to the more complex behaviour previously noted with the congener, LY2033298, at the M2 receptor²⁹. Full methods details are available in the Supplementary Methods.

M2 muscarinic receptor expression and purification. The human M2 muscarinic receptor gene was modified to remove glycosylation sites, and to add an amino-terminal Flag epitope tag and a carboxy-terminal 8 \times His tag. In addition, residues 233–374 of intracellular loop 3 were deleted. This region has previously been shown to be unstructured³⁶ and is not essential for G-protein coupling in the homologous M1 muscarinic receptor³⁷. This construct was expressed in Sf9 insect cells using the BestBac baculovirus system (Expression Systems). Cells were infected at a density of 4×10^6 cells ml⁻¹ and then incubated for two days at 27 °C. Receptor was extracted and purified in the manner described previously for the M3 muscarinic receptor⁸. Briefly, receptor was purified by Ni-NTA chromatography, Flag affinity chromatography and size exclusion chromatography.

Llama immunization samples. M2 receptor was prepared as described above, and bound to iperexo by including it at 10 μ M starting at Flag wash steps and in all

subsequent buffers. Receptor was reconstituted into phospholipid vesicles composed of DOPC (1,2-dioleoyl-sn-glycero-3-phosphocholine, Avanti Polar Lipids) and lipid A in a 10:1 (w:w) ratio, then aliquoted at 1 mg ml⁻¹ receptor concentration and frozen in 100 μ l aliquots before injection.

Yeast display samples. M2 receptor was purified as described above with 1 μ M atropine included in all buffers. Receptor was then labelled with a fivefold molar excess of biotin-NHS ester (Sigma-Aldrich) in buffer containing 25 mM HEPES pH 7.2. After a 30-min incubation at room temperature and a 30-min incubation on ice, unreacted label was quenched with 50 mM Tris pH 8. Directly labelled samples with fluorophore-NHS esters were prepared in a similar manner. Receptor was then desalted into buffer containing either 10 μ M tiotropium, 10 μ M iperexo, or buffer containing no ligand. Receptor eluted in buffer containing no ligand was treated with 50 μ M iperexo mustard (FAUC123; see Supplementary Information for details) for 20 min at room temperature. Samples were then concentrated, aliquoted, and flash frozen with 20% (v/v) glycerol.

Crystallization samples. M2 receptor for crystallization was prepared as described above. When bound to Flag resin, the sample was washed with a mix of dodecyl maltoside buffer (DDM) and buffer containing 0.2% lauryl maltose neopentyl glycol detergent (MNG; Anatrace). These buffers were mixed first in a 1:1 ratio (DDM:MNG buffer), then 1:4 and 1:10 ratios. At each step the 5 ml column was washed with 10 ml of buffer at a 1 ml min⁻¹ flow rate, and all buffers contained 1 μ M atropine. Finally, the column was washed with 10 ml MNG buffer, and then 10 ml of low detergent buffer with agonist (0.01% MNG, 0.001% cholesterol hemisuccinate, 20 mM HEPES pH 7.5, 100 mM NaCl, 10 μ M iperexo). The sample was eluted, mixed with a 1.5-fold stoichiometric excess of Nb9-8 and a second nanobody, NbB4. This nanobody binds to an epitope different from Nb9-8, but was not resolved in the crystal structure. After mixing, the sample was incubated 30 min on ice, then concentrated and purified by size exclusion in low detergent buffer. Eluted protein was concentrated to $A_{280} = 96$, and frozen in liquid nitrogen in 7 μ l aliquots.

Llama immunization. One llama (*Lama glama*) was immunized for 6 weeks with 1 mg receptor in total. Peripheral blood lymphocytes were isolated from the immunized animal to extract total RNA. cDNA was prepared using 50 μ g of total RNA and 2.5 μ g of oligo-dN6 primer. Nanobody open reading frames were amplified as described³⁸.

Post-immune M2 receptor llama nanobody library construction. Nanobody V_{HH} fragments were amplified by PCR using the primers pYalNB80AMPF (5'-C ATTTTCAATTAAGATGCAGTTACTGCTGTTTTCATATTTCTGT ATTGCTAGCGTTTACGAATGGCCAGGTGCAGCTGCAGGAG-3') and pYalNB80AMPR (5'-CCACCAGATCCACCACCACCAAGTCTTCTTCGGA GATAAGCTTTTGTCGGATCCTGAGGAGACGGTGACCTGGGTCCC-3'). The PCR products were then co-transformed with linearized pYal into yeast strain EBY100 as for the Nb80 affinity-maturation library, yielding a library size of 0.6×10^8 transformants.

Selection of M2 G_i-mimetic nanobodies from post-immune M2 llama nanobody library. For the first round of selection, counter-selection was performed against the β_2 receptor to remove yeast clones that bind nonspecifically to membrane proteins or to secondary staining reagents. 1.0×10^9 of induced yeast were washed with PBEM buffer and then stained in 5 ml of PBEM buffer containing 1 μ M biotinylated β_2 receptor liganded with carazolol for 1 h at 4 °C. Yeast were then stained with streptavidin-647 as a secondary reagent and magnetically labelled with anti-647 microbeads (Miltenyi) as described previously¹⁸. Positively labelled yeast were then removed by the use of an LD column (Miltenyi); the cleared flow-through was then used for subsequent selection. Positive selection for clones recognizing the active state of the M2 receptor was performed by staining with 2 μ M biotinylated M2 receptor bound to the agonist iperexo in 5 ml PBEM buffer supplemented with 2 μ M iperexo for 1 h at 4 °C. Yeast were then washed, stained with streptavidin-647, and magnetically labelled with anti-647 microbeads, including 1 μ M iperexo in the PBEM buffer at all steps. Magnetic separation of M2 receptor-binding yeast clones was performed using an LS column (Miltenyi) following the manufacturer's instructions. Magnetically sorted yeast were re-suspended in SDCAA medium and cultured at 30 °C. Rounds 2–4 were selected in a similar manner, counter-selecting against 1 μ M biotinylated β_2 receptor bound to carazolol and positively selecting using 1 μ M biotinylated M2 receptor bound to iperexo. For these rounds, the scale was reduced tenfold to 1×10^8 induced yeast and staining volumes of 0.5 ml.

Conformational selection was performed for rounds 5–9. For rounds 5–8, yeast were stained with 1 μ M biotinylated M2 receptor pre-incubated with the high-affinity antagonist tiotropium for 1 h at 4 °C. Yeast were then fluorescently labelled with either streptavidin-647 or streptavidin-PE, and magnetically labelled with the corresponding anti-647 or anti-PE microbeads (Miltenyi). Depletion of inactive-state binders was carried out using an LS column. The cleared yeast were then positively selected by staining with 0.5 μ M (rounds 5–7) or 0.1 μ M (round 8) M2

receptor pre-bound to iperoxo for 1 h at 4 °C. Yeast were then fluorescently labelled with either streptavidin-PE or streptavidin-647, using a fluorophore distinct from that used in the previous counter-selection step. Magnetic separation of agonist-occupied M2 receptor was performed using an LS column, as for steps 1–4. For round 9, two-colour FACS was performed. Induced yeast were simultaneously stained with 1 µM Alexa647-labelled M2 receptor reacted with iperoxo mustard and 1 µM Alexa488-labelled M2 receptor pre-bound with tiotropium for 1 h at 4 °C. Alexa647 positive/Alexa488 negative yeast were purified using a FACS Jazz cell (BD Biosciences) sorter. Post-sorted yeast were plated onto SDCAA-agar plates and the nanobody-encoding sequences of several colonies were sequenced. Full sequences of clones confirmed to enhance agonist affinity are: Nb9-1, QVQL QESGGGLVQAGGSLRLSQAASGHTFSSARMYWVRQAPGKEREFVAAISRSQFTYSADSVKGRFTISRDIANNITVYLQMNSLPEDTAIYTCYAAAYLDEFYNDYTHYWGLGTQVTVSS; Nb9-8, QVQLQESGGGLVQAGDSLRLSQAASGDFDFFDDYAIWFRQAPGQEREGVSCIDPSDGSTIYADSAKGRFTISSDNAENTVYLQMNSLPEDTAVYVCSAWTLFHSDEYWGQGTQVTVSS; Nb9-20, QVQLQESGGGLVQPEGLSLTACDTSFGTMNYIAIWFQRQAPEKEREGLATISSIDGRITYADSVKGRFTISRDSAKNMYLQMNNLRPEDTAVYVCSAGPDYSYDGYDESEYWGQGTQVTVSS.

Expression of MBP-nanobody fusions in *E. coli*. Nanobody sequences were subcloned into a modified pMalp2x vector (New England Biolabs) containing an N-terminal, 3C protease-cleavable maltose binding protein (MBP) tag and a C-terminal 8×His tag. Plasmids were transformed into BL21(DE3) cells and protein expression induced in Terrific Broth by addition of IPTG to 1 mM at an OD₆₀₀ of 0.8. After 24 h of incubation at 22 °C, cells were collected and periplasmic protein was obtained by osmotic shock. MBP-nanobody fusions were purified by Ni-NTA chromatography and MBP was removed using 3C protease. Cleaved MBP was separated from the 8×His tagged nanobodies by an additional Ni-NTA purification step. The 8×His tag was subsequently removed using carboxypeptidase A.

Expression and purification of G protein. Heterotrimeric G_i was prepared by expression using a single baculovirus for the human G_{α_{i1}} subunit and a second, bicistronic virus for human Gβ1 and Gγ2 subunits. G protein was expressed in HighFive insect cells, and then purified as described previously for G_s (ref. 10). In brief, G protein was extracted with cholate, purified by Ni-NTA chromatography, detergent exchanged into dodecyl maltoside buffer, and then purified by ion exchange and dialysed before use.

M2 receptor radioligand binding assays with G protein and nanobody. M2 receptor was expressed and purified as described above. Receptor was then reconstituted into HDL particles consisting of apolipoprotein A1 and a 3:2 (mol:mol) mixture of the lipids POPC:POPG (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine: 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine and 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-rac-glycerol) respectively, Avanti Polar Lipids). Binding reactions contained 50 fmol functional receptor, 0.6 nM [³H] N-methyl scopolamine (NMS), 100 mM NaCl, 20 mM HEPES pH 7.5, 0.1% BSA, and ligands and nanobodies as indicated. Concentration-dependent effects of nanobodies were measured in the presence of 10 nM iperoxo. All reactions were carried out in a 500 µl volume. For samples containing G protein, purified G_i heterotrimer from insect cells was added to the reactions at a 1,000-fold dilution from a 200 µM stock, resulting in a large stoichiometric excess over receptor and diluting G protein below the detergent CMC to allow incorporation into HDL particles, essentially as described previously³⁹. Reactions were mixed and then incubated for 2 h. Samples were then filtered on a 48-well harvester (Brandel) onto a filter which had been pre-treated with 0.1% polyethylenimine. All measurements were taken by liquid scintillation counting, and experiments were performed at least in triplicate.

Site-directed mutagenesis. A mammalian expression plasmid coding for the human M2 muscarinic receptor (M2R-pcDNA3.1+) was obtained from the Missouri S&T cDNA Resource Center. Mutant M2 receptors were generated by using the QuikChange site-directed mutagenesis kit (Stratagene) according to the manufacturer's instructions. The identity of all mutant M2 receptor constructs was confirmed by DNA sequencing.

Transient expression of receptor constructs in COS-7 cells. Wild-type and mutant M2 receptors were transiently expressed in COS-7 cells grown in 100 mm dishes, as described previously⁴⁰. For functional studies, the various receptor constructs (3 µg each) were co-expressed with a chimaeric G protein α-subunit (G_{qis}; 3 µg plasmid DNA) in which the last five amino acids of G_{α_q} were replaced with the corresponding G_{α_i} sequence⁴¹.

Radioligand binding studies of mutant and wild-type M2 receptors. Acetylcholine bromide was purchased from Sigma. Iperoxo was a gift of Bristol Myers Squibb. [³H]-NMS (85.5 Ci mmol⁻¹) and 3-quinuclidinyl benzilate ([³H]-QNB; 47.4 Ci mmol⁻¹) were from PerkinElmer Life Sciences (Downers Grove). Radioligand binding studies were carried out with membranes prepared from transfected COS-7 cells as described⁴¹. Forty-eight hours after transfection, cells were collected and re-suspended in 25 mM sodium phosphate buffer (pH 7.4) containing 5 mM MgCl₂.

Membrane homogenates were prepared and re-suspended in the same buffer. [³H]-NMS or [³H]-QNB binding reactions were carried out in the presence of 9 µg of membrane protein for 3 h at room temperature (total volume of the incubation mixture: 0.5 ml). In saturation binding studies, six different concentrations of the radioligand were used ([³H]-NMS, 0.3 nM to 10 nM; [³H]-QNB, 0.05 nM to 20 nM). In competition binding assays, membrane homogenates were incubated with ten different concentrations of acetylcholine (13 nM to 1 mM) or iperoxo (0.13 nM to 10 µM) in the presence of a fixed concentration of radioligand (2 nM [³H]-NMS for all receptors except N404Q; 15 nM [³H]-QNB for N404Q and 0.5 nM [³H]-QNB for wild-type M2 receptor). Nonspecific binding was determined in the presence of 10 µM atropine. Reactions were stopped by rapid filtration through GF/C filters. Data were analysed using Prism 4.0 software (GraphPad Software, Inc.).

Calcium mobilization assay. COS-7 cells co-expressing wild-type or mutant M2 receptor and the hybrid G protein, G_{qis} (ref. 41), were incubated with increasing concentrations of agonists (acetylcholine, 5 nM to 50 µM; iperoxo, 50 pM to 0.5 µM), and increases in intracellular calcium levels were determined in 96-well plates using FLIPR technology (Molecular Devices), as described in detail previously^{42,43}. Agonist concentration–response curves were analysed using Prism 4.0 software.

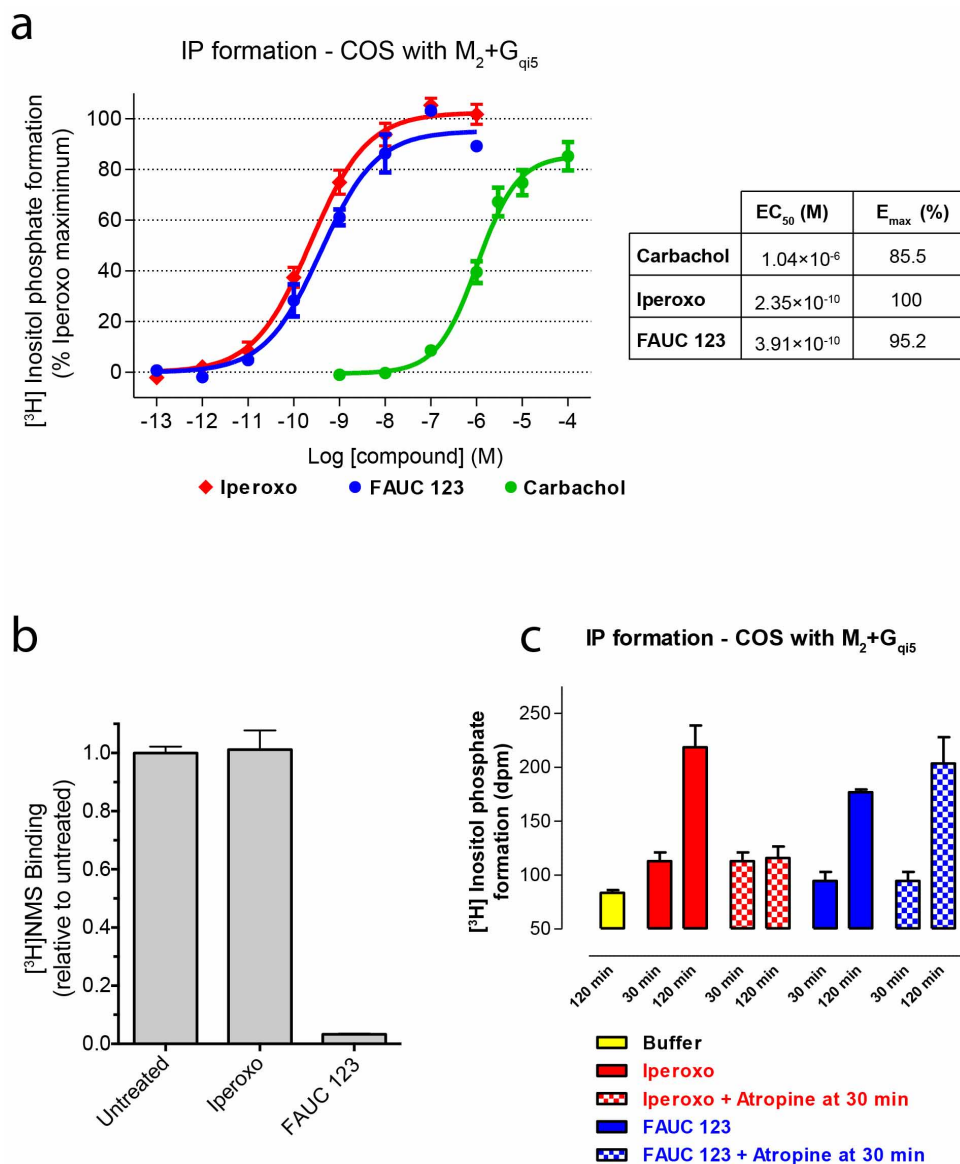
Crystallization. Purified M2 receptor was reconstituted into lipidic cubic phase by mixing with a 1.5-fold excess by mass of 10:1 (w:w) monoolein cholesterol lipid mix. Protein and lipid were loaded into glass syringes (Art Robbins Instruments), and then mixed 100 times by the coupled syringe method⁴⁴. Samples of 30–100 nl in volume were spotted onto 96-well glass plates and overlaid en bloc with 600 nl precipitant solution for each well. Precipitant solution consisted of 10–20% PEG300, 100 mM HEPES pH 7.2–7.9, 1.2% 1,2,3-heptanetriol, and 20–80 mM EDTA pH 8.0. Identical conditions were used to crystallize LY2119620-receptor complexes, except that the overlay precipitant solution was supplemented with 500 µM LY2119620. Crystals grew in 24 h, and reached full size within 2 days. Crystals were then harvested in mesh grid loops (MiTeGen) with 10–50 crystals per loop and stored in liquid nitrogen before use.

Data collection. Grids of crystals were rastered at Advanced Photon Source beamlines 23ID-B and 23ID-D. Initial rastering was performed with an 80 µm by 30 µm beam with fivefold attenuation and 1-s exposure, and regions with strong diffraction were sub-rastered with a 10 µm collimated beam with equivalent X-ray dose. Data collection was similarly performed with a 10 µm beam, but with no attenuation and exposures of typically 1–5 s. An oscillation width of 1–2 degrees was used in each case, and wedges of 5–10 degrees were compiled to create the final data sets.

Data reduction and refinement. Diffraction data were processed in HKL2000⁴⁵, and statistics are summarized in Extended Data Table 1. The structure was solved using molecular replacement with the structure of the inactive M2 receptor (Protein Data Bank accession 3UON) and Nb80 (Protein Data Bank accession 3POG) as search models in Phaser⁴⁶. The resulting structure was iteratively refined in Phenix⁴⁷ and manually rebuilt in Coot⁴⁸. Final refinement statistics are summarized in Extended Data Table 1. Figures were prepared in PyMol (Schrödinger).

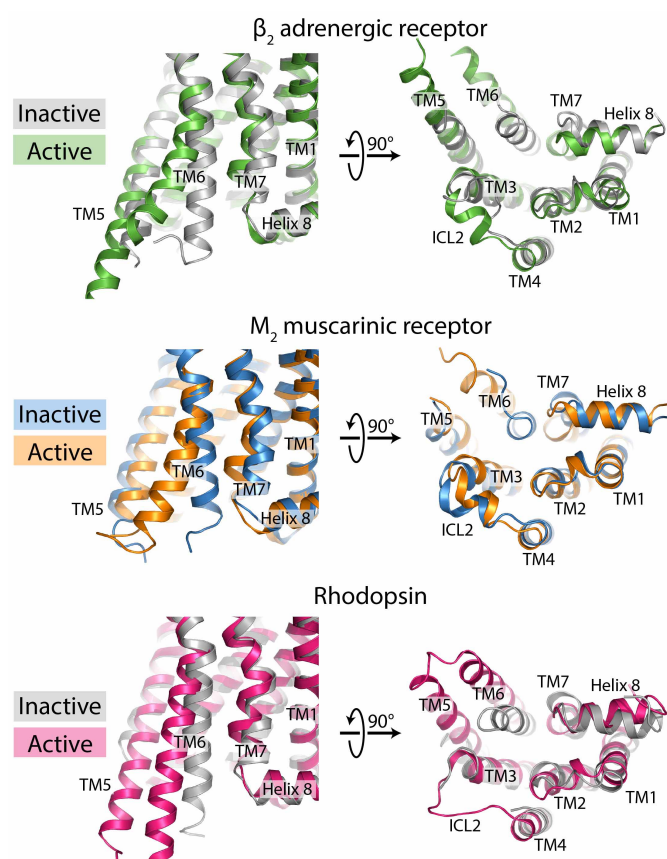
1. Chee, M. J. *et al.* The third intracellular loop stabilizes the inactive state of the neuropeptide Y1 receptor. *J. Biol. Chem.* **283**, 33337–33346 (2008).
2. Broach, J. R. & Thorner, J. High-throughput screening for drug discovery. *Nature* **384**, 14–16 (1996).
3. Ehlert, F. J. Estimation of the affinities of allosteric ligands using radioligand binding and pharmacological null methods. *Mol. Pharmacol.* **33**, 187–194 (1988).
4. Canals, M. *et al.* A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation. *J. Biol. Chem.* **287**, 650–659 (2012).
5. Leach, K., Sexton, P. M. & Christopoulos, A. Allosteric GPCR modulators: taking advantage of permissive receptor pharmacology. *Trends Pharmacol. Sci.* **28**, 382–389 (2007).
6. Ichiyama, S. *et al.* The structure of the third intracellular loop of the muscarinic acetylcholine receptor M2 subtype. *FEBS Lett.* **580**, 23–26 (2006).
7. Shapiro, R. A. & Nathanson, N. M. Deletion analysis of the mouse m1 muscarinic acetylcholine receptor: effects on phosphoinositide metabolism and down-regulation. *Biochemistry* **28**, 8946–8950 (1989).
8. Conrath, K. E. *et al.* β-Lactamase inhibitors derived from single-domain antibody fragments elicited in the camelidae. *Antimicrob. Agents Chemother.* **45**, 2807–2812 (2001).
9. Whorton, M. R. *et al.* A monomeric G protein-coupled receptor isolated in a high-density lipoprotein particle efficiently activates its G protein. *Proc. Natl Acad. Sci. USA* **104**, 7682–7687 (2007).
10. Hu, J. *et al.* Structural basis of G protein-coupled receptor-G protein interactions. *Nature Chem. Biol.* **6**, 541–548 (2010).
11. Liu, J., Conklin, B. R., Blin, N., Yun, J. & Wess, J. Identification of a receptor/G-protein contact site critical for signaling specificity and G-protein activation. *Proc. Natl Acad. Sci. USA* **92**, 11642–11646 (1995).
12. Li, B. *et al.* Rapid identification of functionally critical amino acids in a G protein-coupled receptor. *Nature Methods* **4**, 169–174 (2007).

43. McMillin, S. M., Heusel, M., Liu, T., Costanzi, S. & Wess, J. Structural basis of M3 muscarinic receptor dimer/oligomer formation. *J. Biol. Chem.* **286**, 28584–28598 (2011).
44. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4**, 706–731 (2009).
45. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276 (ed. Carter, C. W.) 307–326 (Academic, 1997).
46. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
47. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
48. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).



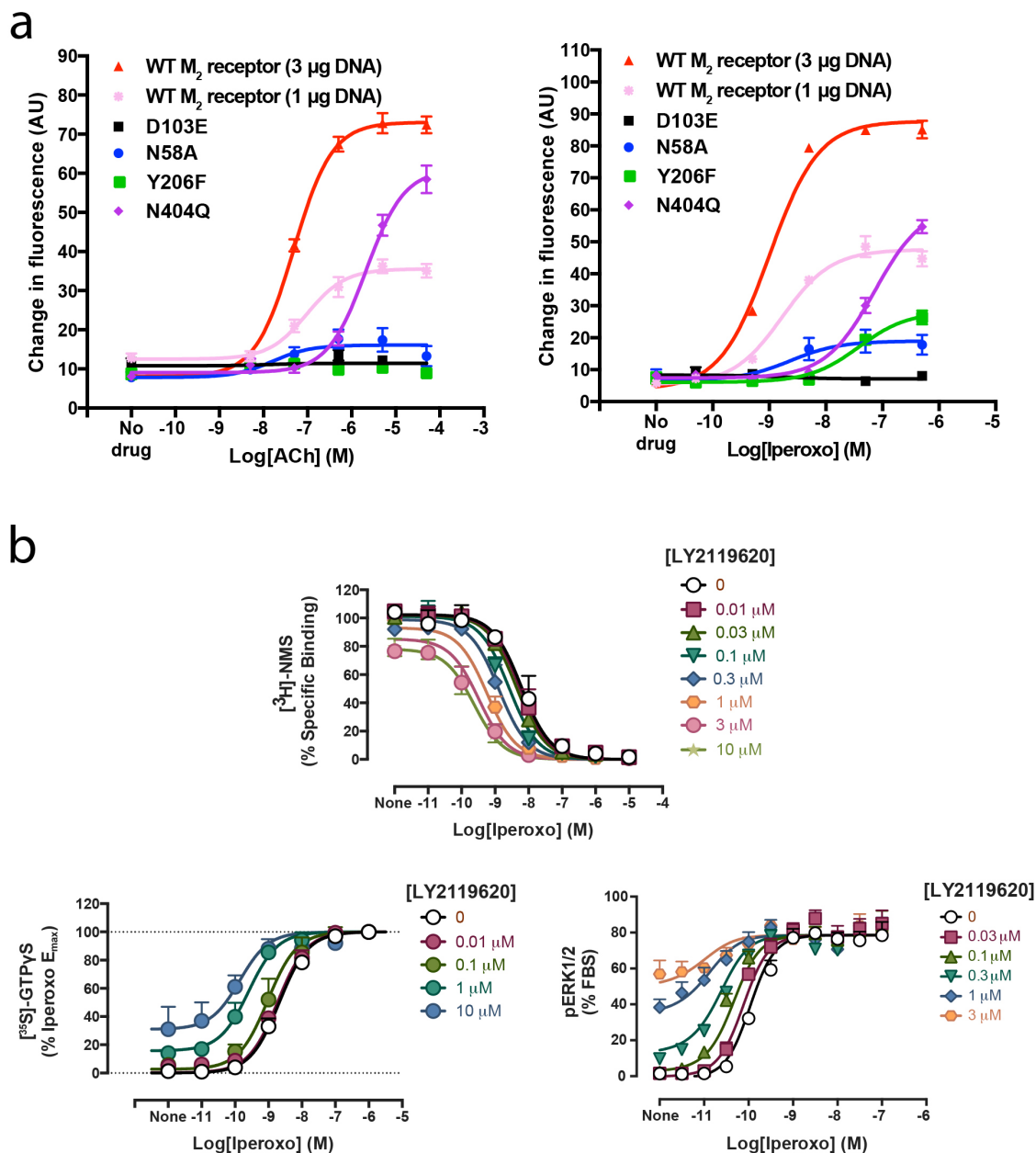
Extended Data Figure 1 | Characterization of FAUC123. **a**, Activation of M₂ receptor by the prototypical muscarinic agonist carbachol, the high-affinity agonist iperoxo, and an irreversible iperoxo analogue (FAUC123) shows that iperoxo and FAUC123 are exceptionally potent full agonists at the M₂ muscarinic receptor. Points indicate mean \pm s.e.m. of three independent measurements, each performed in triplicate. **b**, Sf9 membranes expressing the human M₂ receptor were incubated overnight at 4 °C with either no ligand, 100 μ M iperoxo, or 100 μ M FAUC123. Membranes were then washed three times in buffer without ligand, and incubated with a saturating concentration

(20 nM) of [³H]-NMS. Incubation with iperoxo had no effect on radioligand binding, whereas FAUC123 blocked almost all [³H]-NMS binding sites. Bars indicate mean \pm s.e.m. of three independent measurements. **c**, FAUC123 was tested for its ability to induce M₂ receptor activation after covalent modification. Whereas iperoxo-induced inositol phosphate production was blocked by 1 μ M atropine, FAUC123-induced activation was not susceptible to atropine blockade. Bars indicate mean \pm s.e.m. of three independent measurements.



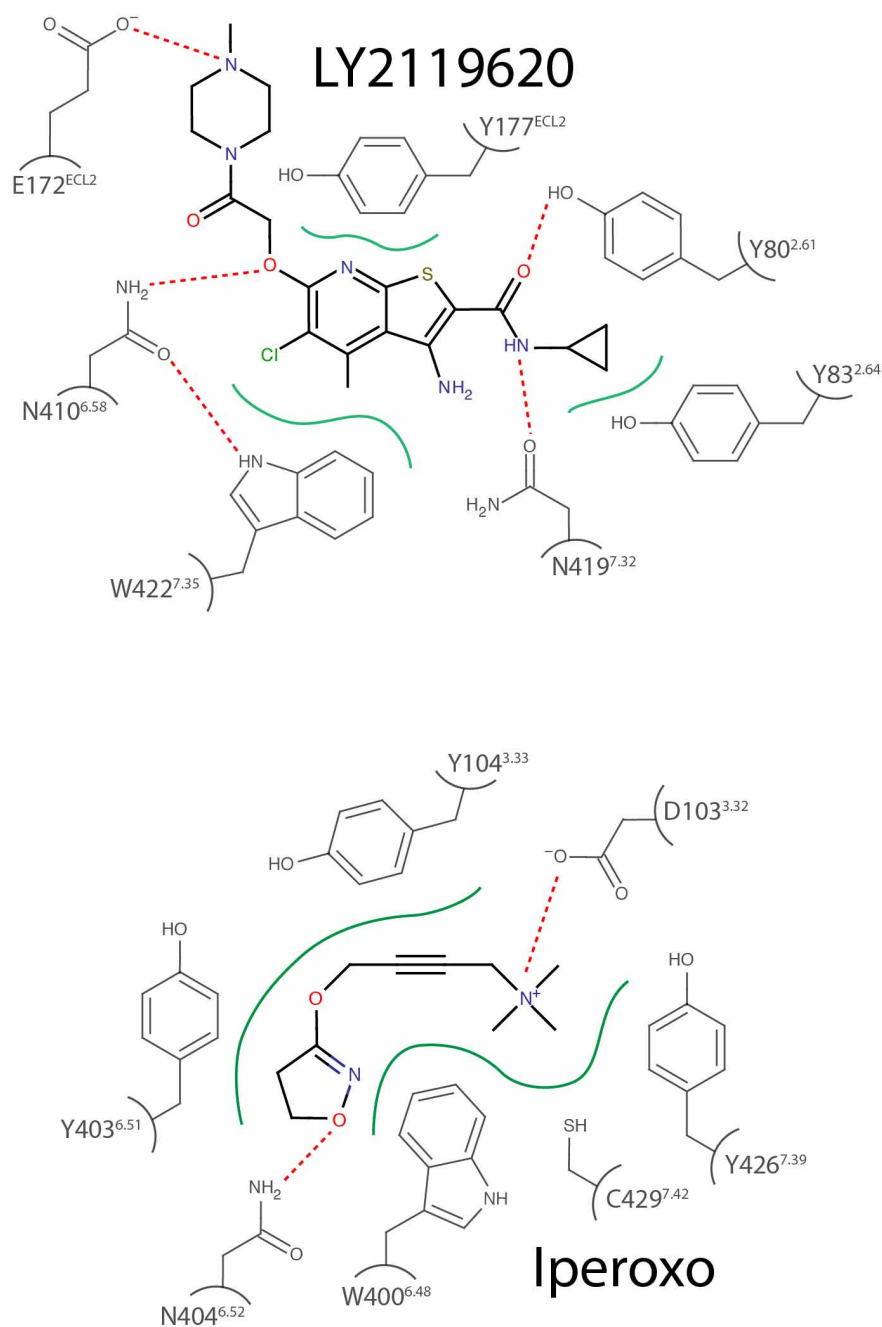
Extended Data Figure 2 | Comparison to other active GPCR structures.

Structures of all activated GPCRs show similarities in conformational changes at the intracellular surface. In each case, the intracellular tip of transmembrane helix 6 (TM6) moves outward on activation, as seen in the view from the intracellular side (right panels). This creates a cavity to which a G protein can bind the receptor.



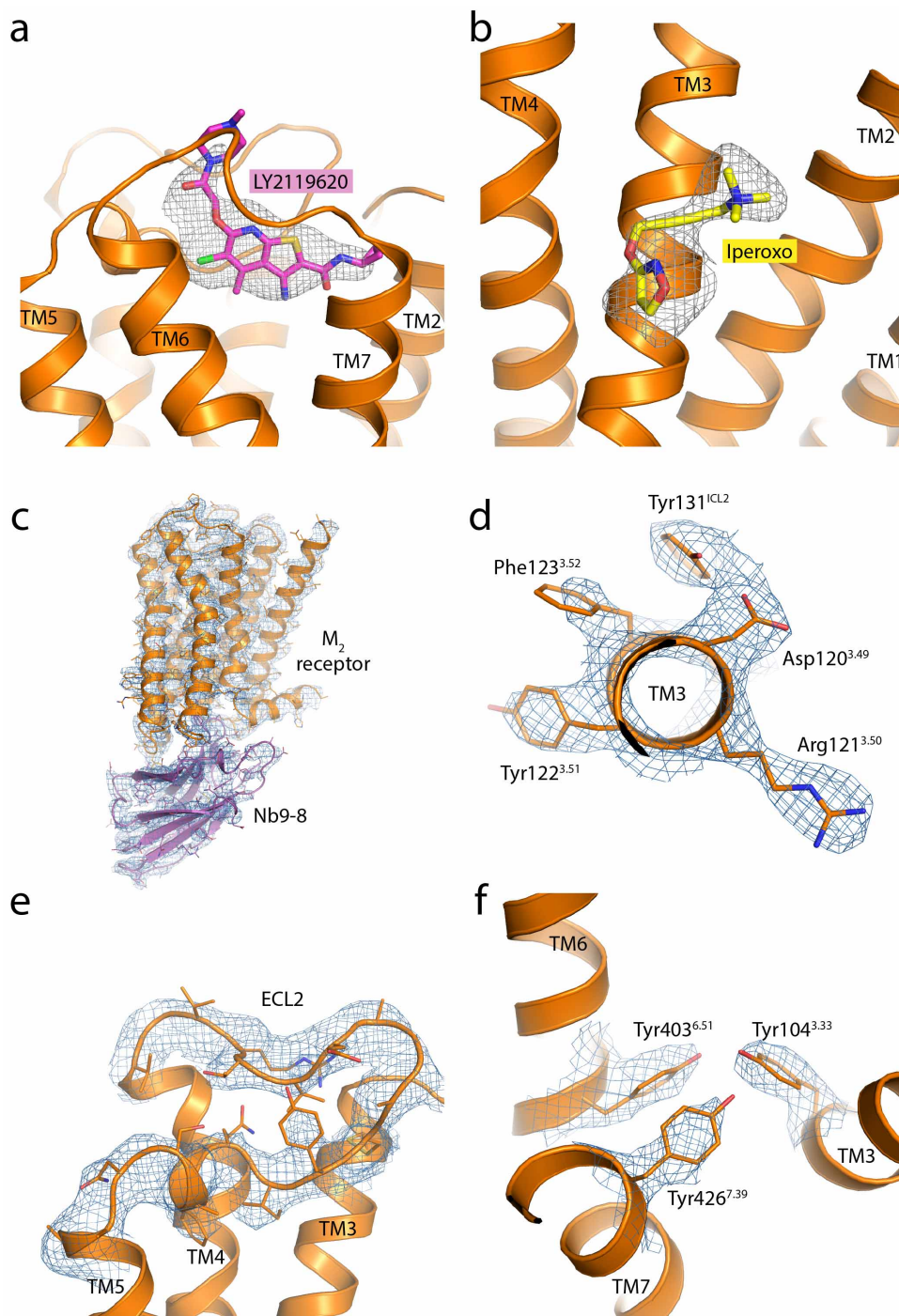
Extended Data Figure 3 | Pharmacology. **a**, Functional properties are shown for M2 receptors in which key residues were mutated. Agonist-induced increases in intracellular calcium levels were monitored via FLIPR using transfected COS-7 cells. Because some mutant receptors (N58A, D103E) were expressed at lower levels than the wild-type (WT) receptor, reference curves were obtained using cells transfected with either 3 µg DNA or 1 µg wild-type receptor DNA. The latter cells showed receptor expression levels comparable to those found with the N58A and D103E mutants (see Extended Data Table 2 for details). Data are given as means \pm s.e.m. of three independent experiments,

each carried out in triplicate. AU, arbitrary units. **b**, The interaction between LY2119620 and iperoxo was measured by radioligand binding and functional assays. LY2119620 enhances the affinity of iperoxo (top graph) and its signalling potency (bottom graphs), and is also able to activate M2 receptor signalling directly as measured by [35 S]GTP γ S and ERK1/2 phosphorylation. Experiments were carried out with CHO cells stably expressing the human M2 receptor, and points are shown as mean \pm s.e.m. of three independent experiments, each carried out in duplicate.



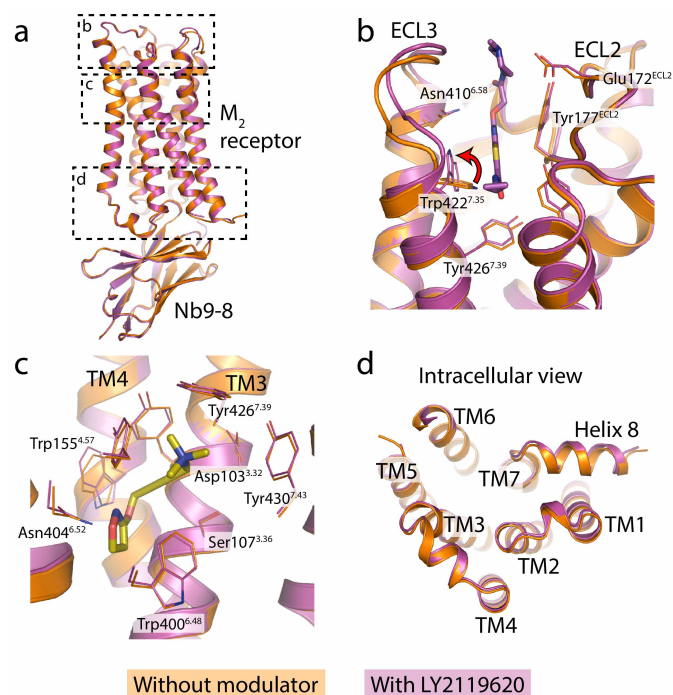
Extended Data Figure 4 | Binding-site diagram. M2 receptor residues interacting with the orthosteric agonist iperoxo and the positive allosteric

modulator LY2119620 are shown. Polar contacts are highlighted as red dotted lines, and hydrophobic contacts are in green solid lines.



Extended Data Figure 5 | Electron density. a, b, $F_o - F_c$ omit maps are shown in grey, contoured at 2.5σ within a 2.5 \AA radius of the indicated ligand.

c-f, $2F_o - F_c$ maps are shown in blue, contoured at 1.5σ within a 2.0 \AA radius of the indicated region.



Extended Data Figure 6 | Comparison of M2 receptor structures with and without LY2119620 bound. Comparison of the structure of active M2 receptor with and without the allosteric modulator LY2119620 reveals that there are few differences outside the extracellular vestibule. The overall structures are compared in **a**. Within the extracellular vestibule, there is a slight contraction in the presence of the modulator, and Trp 422^{7,35} undergoes a change of rotamer (panel **b**, red arrow). The orthosteric ligand-binding site, **c**, and intracellular surface, **d**, show few differences.

Extended Data Table 1 | Data collection and refinement statistics

	M ₂ receptor:Nb9-8 complex	M ₂ receptor:Nb9-8 complex bound to LY2119620
Data collection*		
Number of crystals	17	18
Space group	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁
Unit cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	62.9, 78.1, 163.5	59.0, 77.4, 163.8
<i>α</i> , <i>β</i> , <i>γ</i> (°)	90, 90, 90	90, 90, 90
Resolution (Å)	33 – 3.5 (3.6 – 3.5)	36 – 3.7 (3.8 – 3.7)
R _{merge} (%)	18.8 (74.4)	19.5 (60.5)
<I/σI>	5.8 (1.4)	5.6 (2.1)
CC _{1/2} (%)	99.1 (35.0)	99.0 (54.1)
Completeness (%)	95.9 (83.1)	93.0 (80.1)
Redundancy	4.8	4.8
Refinement		
Number of reflections	10237	7867
R _{work} /R _{free} (%)	24.9 / 29.8	25.0 / 30.1
No. atoms		
Protein	3020	3013
Ligand(s)	14	57
Average B factors (Å ²)		
Receptor	109.7	102.3
Nb6B9	137.4	129.3
Iperoxo	105.5	107.6
LY2119620	-	119.0
RMS deviation from ideality		
Bond length (Å)	0.004	0.004
Bond angles (°)	0.86	0.81
Ramachandran statistics†		
Favored (%)	97.0	95.9
Allowed (%)	3.0	4.1
Outliers (%)	0	0

*Highest shell statistics in parentheses.

†As calculated by Molprobity.

Extended Data Table 2 | Ligand binding properties of mutant M2 receptors

Receptor	$[^3\text{H}]\text{-NMS}$ binding		ACh binding	Iperoxo binding
	K_D	B_{\max}	K_i	K_i
	nM	pmol/mg of protein	μM	μM
WT M ₂	1.48 ± 0.31	1.79 ± 0.17	2.74 ± 0.13	0.0073 ± 0.0006
WT M ₂ (1 μg DNA)	1.40 ± 0.02	0.60 ± 0.14		
N58A ^{2,39}	1.15 ± 0.05	0.62 ± 0.14	0.84 ± 0.21	0.0053 ± 0.0012
D103E ^{3,32}	2.57 ± 0.58	0.51 ± 0.10	327 ± 91	2.76 ± 0.82
Y206F ^{5,58}	1.67 ± 0.26	1.87 ± 0.26	31.8 ± 0.71	0.52 ± 0.20
N404Q ^{6,52}	N.D.*			

Receptor	$[^3\text{H}]\text{-QNB}$ binding		ACh binding	Iperoxo binding
	K_D	B_{\max}	K_i	K_i
	nM	pmol/mg of protein	μM	μM
WT M ₂	0.058 ± 0.015	1.63 ± 0.19	1.25 ± 0.07	0.0130 ± 0.0070
N404Q ^{6,52}	9.47 ± 2.22	1.24 ± 0.12	34.3 ± 10.3	1.70 ± 0.29

Radioligand binding studies were carried out with membranes prepared from COS-7 cells transiently expressing the indicated mutant M2 receptor constructs. The wild-type M2 receptor was expressed at two different densities to allow for a more straightforward interpretation of the functional data shown in Extended Data Fig. 3 (see Methods for details). Acetylcholine and iperoxo binding affinities (K_i) were determined in radioligand competition binding assays as indicated. Acetylcholine and iperoxo binding affinities (K_i) were determined in $[^3\text{H}]\text{-QNB}$ competition binding assays for the N404Q^{6,52} mutant, which did not bind $[^3\text{H}]\text{-NMS}$ with sufficient affinity. Data are given as means \pm s.e.m. from two or three independent experiments, each performed in duplicate.

* No detectable $[^3\text{H}]\text{-NMS}$ binding activity.

Extended Data Table 3 | Pharmacological characterization of LY2119620**a**

Parameter	Value
pK_B^*	5.77 ± 0.10
pK_i^\dagger	8.51 ± 0.04
$\text{Log}\alpha^\ddagger$	1.40 ± 0.09 ($\alpha = 25$)
$\text{Log}\alpha'^\S$	-0.26 ± 0.03 ($\alpha' = 0.6$)

b

Parameter	$[^3\text{S}]\text{GTP}\gamma\text{S}$	ERK1/2
pK_B^*	5.73 ± 0.11	5.84 ± 0.18
$\text{log}\alpha\beta^{\parallel}$	1.42 ± 0.09 ($\alpha\beta = 26$)	1.30 ± 0.20 ($\alpha\beta = 20$)
$\text{log}\tau_B^\P$	-0.28 ± 0.05	0.33 ± 0.09

a, Allosteric ternary complex model binding parameters for the interaction between LY2119620, iperoxo and $[^3\text{H}]\text{-NMS}$ at the human M2 receptor. **b**, Operational model parameters for the functional allosteric interaction between iperoxo and LY2119620 at the human M2 receptor. Estimated parameter values represent the mean \pm s.e.m. of three experiments performed in duplicate.

* Negative logarithm of the equilibrium dissociation constant of LY2119620.

\dagger Negative logarithm of the equilibrium dissociation constant of iperoxo.

\ddagger Logarithm of the binding cooperativity factor between LY2119620 and iperoxo (antilogarithm shown in parentheses).

\S Logarithm of the binding cooperativity factor between LY2119620 and $[^3\text{H}]\text{-NMS}$ (antilogarithm shown in parentheses).

\parallel Logarithm of the product of the binding cooperativity (α) and activation modulation (β) factors between iperoxo and LY2119620. Antilogarithm shown in parentheses.

\P Logarithm of the operational efficacy parameter of LY2119620 as an allosteric agonist.

Structure of the TRPV1 ion channel determined by electron cryo-microscopy

Maofu Liao^{1*}, Erhu Cao^{2*}, David Julius² & Yifan Cheng¹

Transient receptor potential (TRP) channels are sensors for a wide range of cellular and environmental signals, but elucidating how these channels respond to physical and chemical stimuli has been hampered by a lack of detailed structural information. Here we exploit advances in electron cryo-microscopy to determine the structure of a mammalian TRP channel, TRPV1, at 3.4 Å resolution, breaking the side-chain resolution barrier for membrane proteins without crystallization. Like voltage-gated channels, TRPV1 exhibits four-fold symmetry around a central ion pathway formed by trans-membrane segments 5–6 (S5–S6) and the intervening pore loop, which is flanked by S1–S4 voltage-sensor-like domains. TRPV1 has a wide extracellular ‘mouth’ with a short selectivity filter. The conserved ‘TRP domain’ interacts with the S4–S5 linker, consistent with its contribution to allosteric modulation. Subunit organization is facilitated by interactions among cytoplasmic domains, including amino-terminal ankyrin repeats. These observations provide a structural blueprint for understanding unique aspects of TRP channel function.

TRP channels represent a large and diverse family of non-selective cation channels that respond to a wide range of chemical and physical stimuli^{1,2}. Genetic and pharmacological studies highlight the importance of TRP channels in numerous biological processes ranging from calcium adsorption to sensory transduction³, and thus elucidating how these channels respond to physiological stimuli or drugs is relevant to understanding disorders affecting every major organ system in the body. TRP channels are believed to resemble voltage-gated potassium (K_V) or sodium (Na_V) channels in overall transmembrane topology and subunit organization¹. However, aside from the fact that some TRP channels exhibit mild voltage sensitivity⁴, they otherwise share little in common with voltage-gated ion channels (VGICs) with regards to pharmacological and biophysical properties. Whereas crystal structures have been determined for VGICs^{5–8}, this has not been accomplished for any member of the TRP channel family. Thus, it is currently unknown whether, or to what extent, TRP and K_V or Na_V channels share a common structural core, or how structural similarities and differences account for unique functionalities.

TRPV1 is the receptor for capsaicin, the pungent agent from chili peppers that elicits burning pain⁹. It is also the founding member of a subfamily of thermosensitive TRP channels that enable somatosensory neurons, or other cell types, to detect changes in ambient temperature. TRPV1 is activated by noxious heat and modulated by inflammatory agents, such as extracellular protons and bioactive lipids, which contribute to pain hypersensitivity^{9–13}. TRPV1 is arguably the best-characterized member of the vertebrate TRP family; its widely validated role in pain physiology and the availability of well-characterized pharmacological agents make it a ‘poster child’ for elucidating basic principles underlying TRP channel function and structure. Moreover, TRPV1 and other somatosensory TRP channels are considered important targets for analgesic drugs^{12,13}, providing further impetus to determine a structure for any member of this extended protein family.

In recent years, single-particle electron cryomicroscopy (cryo-EM) has enabled three-dimensional (3D) reconstruction of large protein complexes to near-atomic resolution^{14–17}, but analysis of small membrane

proteins, such as TRP channels^{18–20}, remained at low resolution. Here, we exploit a newly developed direct electron detector and new image-processing algorithms to correct motion-induced image blurring and improve signal and contrast of single-particle cryo-EM images^{14,15}. With these tools, we determine the structure of TRPV1 at 3.4 Å resolution without crystallization. Thus, in addition to revealing a TRP channel structure, we showcase single-particle cryo-EM as a powerful and transformative advance in the structural analysis of membrane proteins.

General architecture of TRPV1

We first identified a rat TRPV1 deletion mutant with enhanced biochemical stability (Extended Data Figs 1 and 2). When expressed in mammalian cells or oocytes this modified subunit responded to numerous TRPV1 stimuli, including capsaicin, resiniferatoxin, extracellular protons, spider toxins, and heat, and like the wild-type channel showed relatively high permeability for calcium^{21,22}. Thus, biophysical and structural information gleaned from this minimal TRPV1 construct should accurately reflect properties of the wild-type channel.

We initiated structural analysis using negative-stain EM (Extended Data Fig. 3) followed by cryo-EM 3D reconstruction of images recorded at 200 kV with a phosphor scintillator-based complementary metal-oxide-semiconductor (CMOS) camera (Extended Data Figs 4 and 5). We then collected a cryo-EM data set at 300 kV using a direct detection camera following newly implemented procedures of dose fractionation and motion correction¹⁵. Thin rings were visible to ~ 3 Å in the Fourier power spectrum of almost every motion-corrected image (Fig. 1a, b and Extended Data Fig. 6). Two-dimensional (2D) class averages of particle images (Fig. 1c and Extended Data Fig. 7) showed greater detail than those calculated from CMOS camera images (Extended Data Fig. 4e, f), demonstrating considerable improvement of data quality. The final 3D reconstruction of TRPV1 has an overall resolution of 3.4 Å, using gold-standard Fourier shell correlation (FSC) = 0.143 criteria (Fig. 1d–g and Extended Data Fig. 8)²³.

Side-chain densities of most residues within transmembrane helices S1–S6 were clearly resolved, as well as the TRP domain following

¹Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California, San Francisco, California 94158-2517, USA. ²Department of Physiology, University of California, San Francisco, California 94158-2517, USA.

*These authors contributed equally to this work.

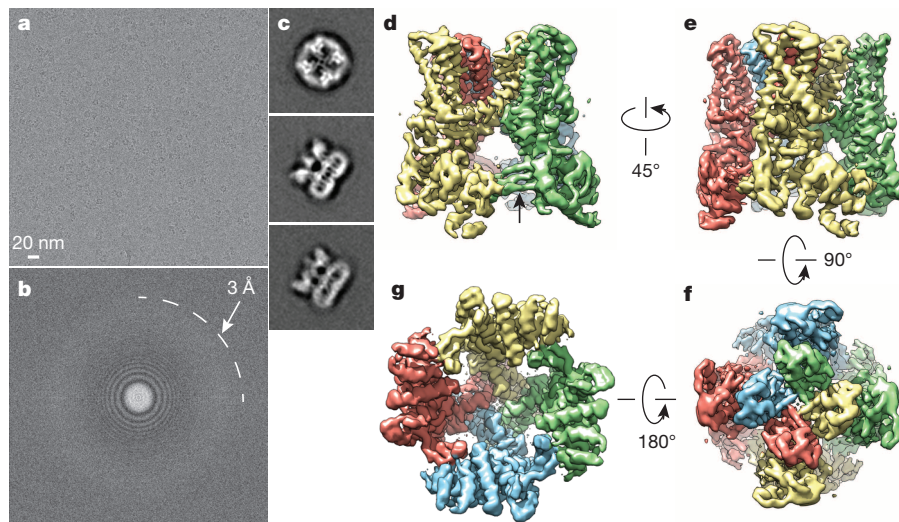


Figure 1 | 3D reconstruction of TRPV1 determined by single-particle cryo-EM. **a**, Representative electron micrograph of TRPV1 protein embedded in a thin layer of vitreous ice recorded at a defocus of 1.7 μm . **b**, Fourier transform of micrograph shown in **a**, with Thon rings extending to nearly 3 \AA . **c**, Enlarged views of three representative 2D class averages show fine features of

S6 and most ($\sim 80\%$) residues in the N-terminal segment connecting the ankyrin repeat domain (ARD) to S1 (Extended Data Fig. 9a–d). In other cytoplasmic regions, densities were less well resolved, but anti-parallel β -strands were well separated, indicating that cytoplasmic domains were resolved to ≤ 4.8 \AA (Fig. 1d, arrow). Densities for the first two ankyrin repeats were absent in our 3D density map (Fig. 1 and Extended Data Figs 5 and 8), even though the entire ARD was present in the expressed protein. Nonetheless, the crystal structure of isolated TRPV1 ankyrin repeats²⁴ could be nicely docked to the density map as a rigid body without modification, as confirmed by bulky-side-chain densities (Extended Data Fig. 9e). Finally, we constructed an atomic model of TRPV1 in which residues L360 to A719 were built *de novo*, and L111 to E359 were docked from the ankyrin repeat crystal structure (Fig. 2). In the carboxy-terminal domain, we could trace the main chain, but side-chain densities were not sufficient for *de novo* model building.

Our TRPV1 structure demonstrates that, despite the rather low ($<20\%$) sequence similarity between TRPs and VGICs, these two channel families share a similar tetrameric architecture in which subunits are arranged in four-fold symmetry around a central ion permeation path. Each subunit consists of six transmembrane α -helices (S1–S6) that span the lipid bilayer (Fig. 2 and Extended Data Fig. 8), plus a re-entrant loop with a pore helix located between S5 and S6 that together assume an ‘inverted teepee’ arrangement resembling that of VGICs (Fig. 2c).

Domains unique to TRP channels are immediately evident. First, we see four of six known ankyrin repeats at the N terminus, arranged as described previously²⁴, followed by a linker (P360–V415) that is conserved among TRPV subtypes and which connects the ARD to the pre-S1 helix (Fig. 3). This linker forms a tightly packed domain containing β -strand and α -helical elements, and is sandwiched between the sixth ankyrin repeat on one side and the pre-S1 helix and TRP domain on the other. Interestingly, an anti-parallel β -sheet from the linker region and a β -strand from the C terminus make contact with two ankyrin repeats from an adjacent subunit (Figs 1d, 2a and 3b), an interaction that probably has a role in channel assembly, reminiscent of the T1 domain in K_V channels. Another unique feature is the extended and kinked interfacial helix following S6, which corresponds to the signature ‘TRP domain’ found in many TRP channels². This domain is strategically located to interact with both the S4–S5 linker and pre-S1 helix (E416–R428), probably providing a physical substrate through which stimuli allosterically affect pore conformation (Fig. 3).

tetrameric channel complex. **d–g**, 3D density map of TRPV1 channel filtered to a resolution of 3.4 \AA (scaled to atomic structure) with each subunit colour-coded. Four different views of the channel are shown, from side (**d**, **e**), top (**f**) and bottom (**g**). The arrow in panel **d** indicates β -sheet structure in the cytosolic domain of TRPV1.

The transmembrane core

In VGICs, the transmembrane core is organized into two distinct clusters that include the central ion-conducting pore formed by tetrameric assembly of S5–P–S6 domains, and surrounding S1–S4 voltage-sensing domains^{25,26}. Each S1–S4 domain associates with the S5–P–S6 region of an adjacent subunit, forming a pinwheel-like structure through this ‘domain swap’ organization. Notably, we find that TRPV1 adopts this same overall configuration (Fig. 2d), despite minimal sequence and functional similarity to VGICs^{1,4}. When pore regions (S5–P–S6) of three channels (TRPV1, K_V 1.2–2.1 chimera and Na_V Ab) are aligned, the relative positions of S1–S4 domains differ, even though the domains themselves show substantial overlap (Fig. 4a, b), perhaps reflecting intrinsic structural differences or distinct physiological states (for example, open, closed or desensitized). Within each TRPV1 subunit, the S1–S4 and S5–P–S6 units are connected by a helical S4–S5 linker that runs parallel to the membrane, again reminiscent of VGICs (Figs 3 and 4c and Extended Data Fig. 10a).

In VGICs, S3 and S4 helices engage in extensive side-chain interactions, forming a charge-transfer centre that facilitates movement of S4 in response to changes in membrane potential^{5,7}. Interestingly, the equivalent region of TRPV1 lacks charged side chains and is, instead, replete with aromatic residues (Fig. 4d and Extended Data Fig. 10b). The importance of this aromatic cluster is underscored by studies demonstrating that replacement of these amino acids (Y441, Y444, Y554 and Y555) with small non-aromatic residues produces non-functional channels^{27,28}. Such hydrophobic packing might impart rigidity to the S1–S4 domain, allowing it to serve as a relatively stationary anchor upon which the S4–S5 linker moves to facilitate TRPV1 gating. Indeed, as discussed in the accompanying study²⁹, the entire S1–S4 domain remains static during channel activation, but provides an external surface for the binding of lipophilic ligands, such as capsaicin and resiniferatoxin. Thus, unique biochemical interactions within this structurally conserved S1–S4 domain probably reflect differential sensitivities of VGICs and TRPs to specific physiologic stimuli (that is, voltage versus lipid metabolites).

The TRP domain and S4–S5 linker

The TRP domain, a 23–25-amino-acid-long region located just after S6, is found in many TRP family members and has been proposed to engage in subunit assembly or allosteric modulation of channel gating^{1,2,30}. We find that the TRP domain assumes a α -helical structure that runs parallel to the inner leaflet of the membrane by virtue of a sharp bend

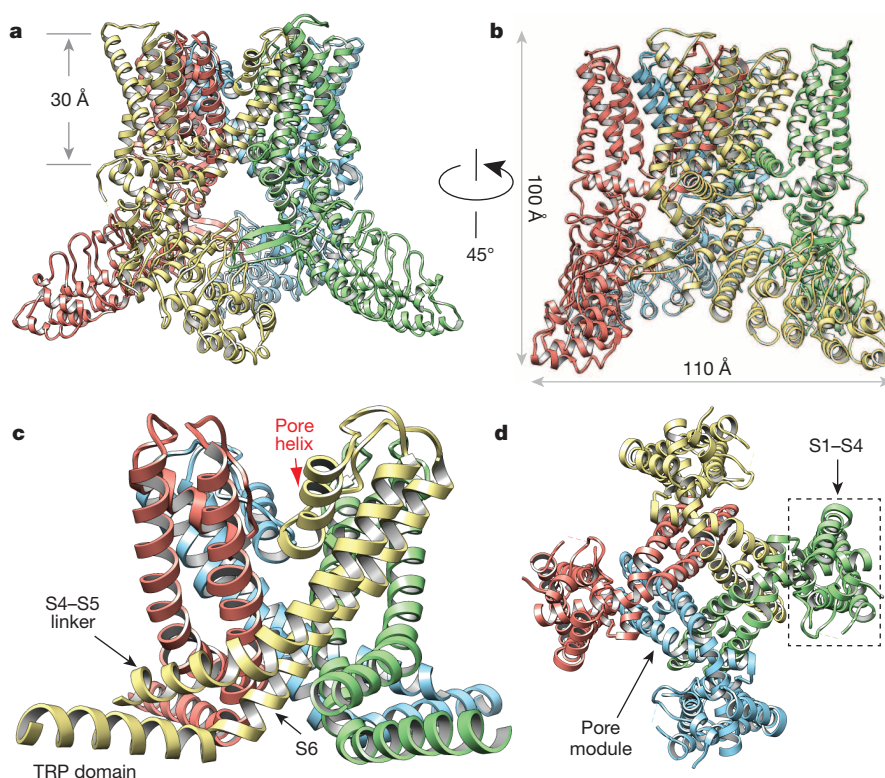


Figure 2 | TRPV1 and VGICs share similar four-fold symmetric architecture. **a–d**, Ribbon diagram of TRPV1 atomic model with each of the four identical subunits colour-coded, showing views from side (**a** and **b**). The dimensions of the channel and the length of the membrane-spanning helices are indicated. The dimensions of the TRPV1 tetramer are $100 \text{ Å} \times 110 \text{ Å} \times 110 \text{ Å}$, as compared with $135 \text{ Å} \times 95 \text{ Å} \times 95 \text{ Å}$ for the rat

Kv1.2 potassium channel⁵. **c**, Ribbon diagram focusing in on side view of S5–P–S6 pore with TRP domains. **d**, Bottom view focusing on transmembrane core, including S1–S4, S5–P–S6 and TRP domains. Note that S1–S4 domains flank and interact with S5–P–S6 pore modules from adjacent subunit, reminiscent of VGIC architecture.

after S6 (Fig. 2c). This α -helix encompasses the first two-thirds of the TRP domain, after which the structure transitions to a random coil (Fig. 3b). An invariant tryptophan (W697) near the middle of the TRP domain forms a hydrogen bond with the main-chain carbonyl oxygen of F559 at the beginning of the S4–S5 linker (Extended Data Fig. 10c). Interestingly, gain-of-function mutations at this equivalent tryptophan in TRPV3 underlie a congenital disorder, Olmsted syndrome, in humans³¹. Moreover, charged side chains within the TRP domain are located on the side of the helix facing the cytoplasm, where they interact with the

pre-S1 helix through hydrogen bonding and salt bridging (Fig. 4e and Extended Data Fig. 10c, d). Taken together, these observations are consistent with a role for the TRP helix as a point of structural integration that facilitates allosteric coupling between channel domains.

In VGICs, the S4–S5 linker has a critical role in coupling the movement of the S1–S4 voltage sensor to gating of the pore^{5–8,32}. Gain-of-function mutations within the S4–S5 linker or S5 helix of TRPV4 enhance basal open probability and result in skeletal dysplasia syndromes³³. Interestingly, a cation– π interaction can be seen between residues at

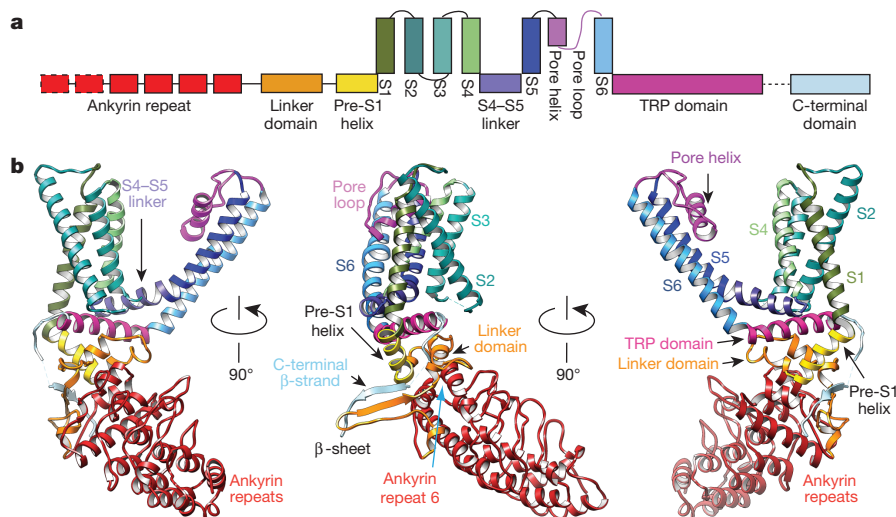


Figure 3 | Structural details of a single TRPV1 subunit. **a**, Linear diagram depicting major structural domains in a TRPV1 subunit, colour coded to match ribbon diagrams below. Dashed boxes denote regions for which density was not

observed (first two ankyrin repeats) or where specific residues could not be definitively assigned (C-terminal β -strand). **b**, Ribbon diagrams showing three different views of a TRPV1 monomer denoting specific domains.

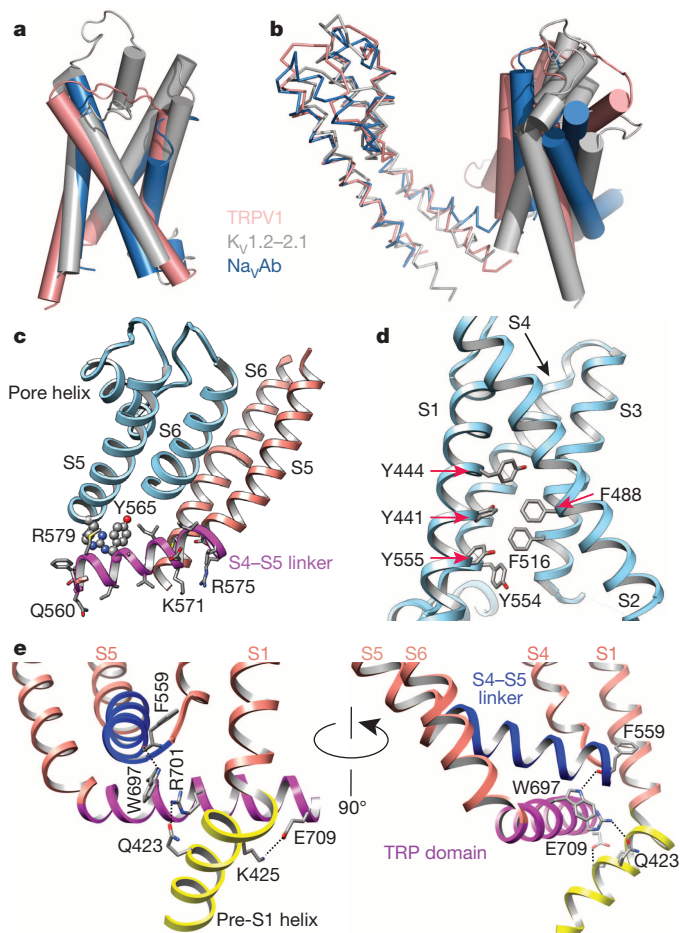


Figure 4 | Unique structural features of TRPV1. **a**, Alignment of S1–S4 transmembrane domains from TRPV1 (salmon), $K_v1.2-2.1$ chimaera (PDB 2R9R; grey) and Na_vAb (PDB 3RVY; blue) show substantial overlap. **b**, When S5–P–S6 pore regions are aligned, the S1–S4 domains show differential relative orientations. **c**, The S4–S5 linker is an amphipathic α -helix whose charged surface faces the cytosol. Potential cation– π interactions between Y565 in the S4–S5 linker and R579 from S5 of the adjacent subunit are highlighted. Mutations of cognate residues in TRPV4 render the channel constitutively active and cause skeletal dysplasia. **d**, Aromatic side chains from S1, S3 and S4 helices create a hydrophobic interior in the S1–S4 domains, in contrast to the charged environment observed in VGICs. **e**, Two different views highlight interactions between TRP domain and S4–S5 linker and pre-S1 helix. Interactions (that is, hydrogen bonds and salt bridge) are indicated by dashed lines.

equivalent positions in TRPV1, namely Y565 in the S4–S5 linker of one subunit and R579 in S5 of a neighbouring subunit (Fig. 4c and Extended Data Fig. 10a). Thus, in addition to covalent intra-subunit interaction between the S4–S5 linker and S5–P–S6 pore region, functional coupling may also be achieved through weak trans-subunit interactions.

The pore and ion permeation pathway

The outer pore region of TRPV1 is wide open compared to K_v channels, with a broad funnel-like structure that probably enhances accessibility to both small and large pharmacophores (Fig. 5a, b). Indeed, as described in the accompanying study²⁹, the uppermost regions of the pore loop proximal to S5 and S6 form a binding site for spider toxins. Further down the central canal, we see a short selectivity filter (⁶⁴³GMGD⁶⁴⁶) in which backbone carbonyls or side chains point into the central pathway (Fig. 5a–e). TRPV1 and related subtypes are highly calcium permeable^{9,34} and molecular-modelling studies predict a selectivity filter diameter of ~ 6 Å^{35,36}. In our structure, we see a restriction point of 4.6 Å between diagonally opposed carbonyl oxygens at G643 (Fig. 5a–c, e), suggesting

that the selectivity filter is in a non-conducting state, reminiscent of inactivated bacterial Na_vRh channels⁸. Indeed, as we show in the accompanying study²⁹, the selectivity filter diameter increases substantially when TRPV1 is in its activated state. The overall arrangement that we see for TRPV1—namely, a wide outer pore with a short selectivity filter—more closely resembles that of bacterial Na_v channels than K_v channels (Extended Data Fig. 10e, f)^{5–8}. However, TRPV1 lacks hydrogen bonding within and between adjacent pore helices, in contrast to Na_v channels, in which such elaborate interactions are believed to impart rigidity to the selectivity filter⁷. This more flexible architecture, together with a short selectivity filter, may underlie the phenomenon of pore dilation seen in several TRP channels, including TRPV1, whereby prolonged or repetitive activation renders the channel permeable to large organic cations^{21,37}. Interestingly, when compared with TRPV1–4, TRPV5 and TRPV6 subtypes bear distinct sequences at the selectivity filter (that is, T–V/I–I–D versus G–M/L–G–D/E), as well as the pore helix, suggesting a different architecture of the outer pore domain (Extended Data Fig. 2). This divergence may underlie the fact that TRPV5 and TRPV6 exhibit more pronounced selectivity for calcium over monovalent cations (permeability ($P_{Ca^{2+}}/P_{Na^+}$) > 100) compared to other TRPV subtypes ($P_{Ca^{2+}}/P_{Na^+} \approx 3–10$) (ref. 34).

Continuing down the pore, we see a constriction site in which I679 in S6 helices from each subunit come together to form a hydrophobic seal measuring 5.3 Å between side chains, thus constituting the most constricted point in the lower gate (Fig. 5a–c, g). Modelling studies, together with substituted cysteine accessibility analysis, have previously implicated I679, Y671 and L681 as candidates for this position^{35,38}. Our structure indicates that side chains of Y671 are located ~ 10 Å above the narrowest point formed by I679 (Fig. 5a, f), and that the side chain of L681 points away from the ion conduction pathway, making Y671 and L681 unlikely major contributors to the lower gate. Of note, TRPV subtypes exhibit high sequence similarity throughout the S6 helix, and all contain an isoleucine at the position equivalent to I679 in TRPV1 (Extended Data Fig. 2), suggesting a similar architecture in the lower gate. Whereas the narrowest constriction site in bacterial KcsA and Na_vAb channels is located at the apex of the inverted teepee^{7,39}, the I679 restriction point is higher in the TRPV1 permeation pathway, sitting ~ 10 Å above the S6 bundle crossing. We believe that the structure shown here represents TRPV1 in the closed state, a conclusion that is validated by comparison with structures representing agonist-bound states of the channel described in the accompanying study²⁹.

Ankyrin repeats form an assembly domain

A characteristic feature of many TRP channels is the presence of ankyrin repeats within the cytoplasmic N terminus⁴⁰. In our structure, residues in the atypically long finger 3 and inner helices of ankyrin repeats 3 and 4 from one subunit interact with a three-stranded antiparallel β -sheet formed by the ARD–S1 linker region (K368–D383) and C terminus of an adjacent subunit, packing the cytosolic part of the channel together (Fig. 6). Moreover, the β -sheet structure tethers cytoplasmic N- and C-terminal domains together within the same subunit, reminiscent of G-protein-coupled inwardly rectifying potassium channels, in which N- and C termini interact through a short parallel β -sheet^{41,42}. Interestingly, two residues (G375 and P376) that produce the sharp turn connecting the N-terminal β -strands are invariant among all TRPV subtypes (Extended Data Fig. 2), supporting the proposed functional importance of the β -sheet in subunit structure and assembly.

In the TRPV1 ARD crystal structure, ATP is bound within the concave surface formed by ankyrin repeats 1–3, and has been proposed to both stabilize the ARD fold and regulate Ca^{2+} -calmodulin-dependent desensitization²⁴. Moreover, a number of mutations that render TRPV1 or other TRPV channels constitutively active lie in the vicinity of the proposed ATP binding site^{22,43}, further indicating that the ARD is an important locus for channel modulation. Our structural data suggest that modulation may also involve perturbation of subunit–subunit interactions. At the same time, the ARD surface forms an extensive cytoplasmic

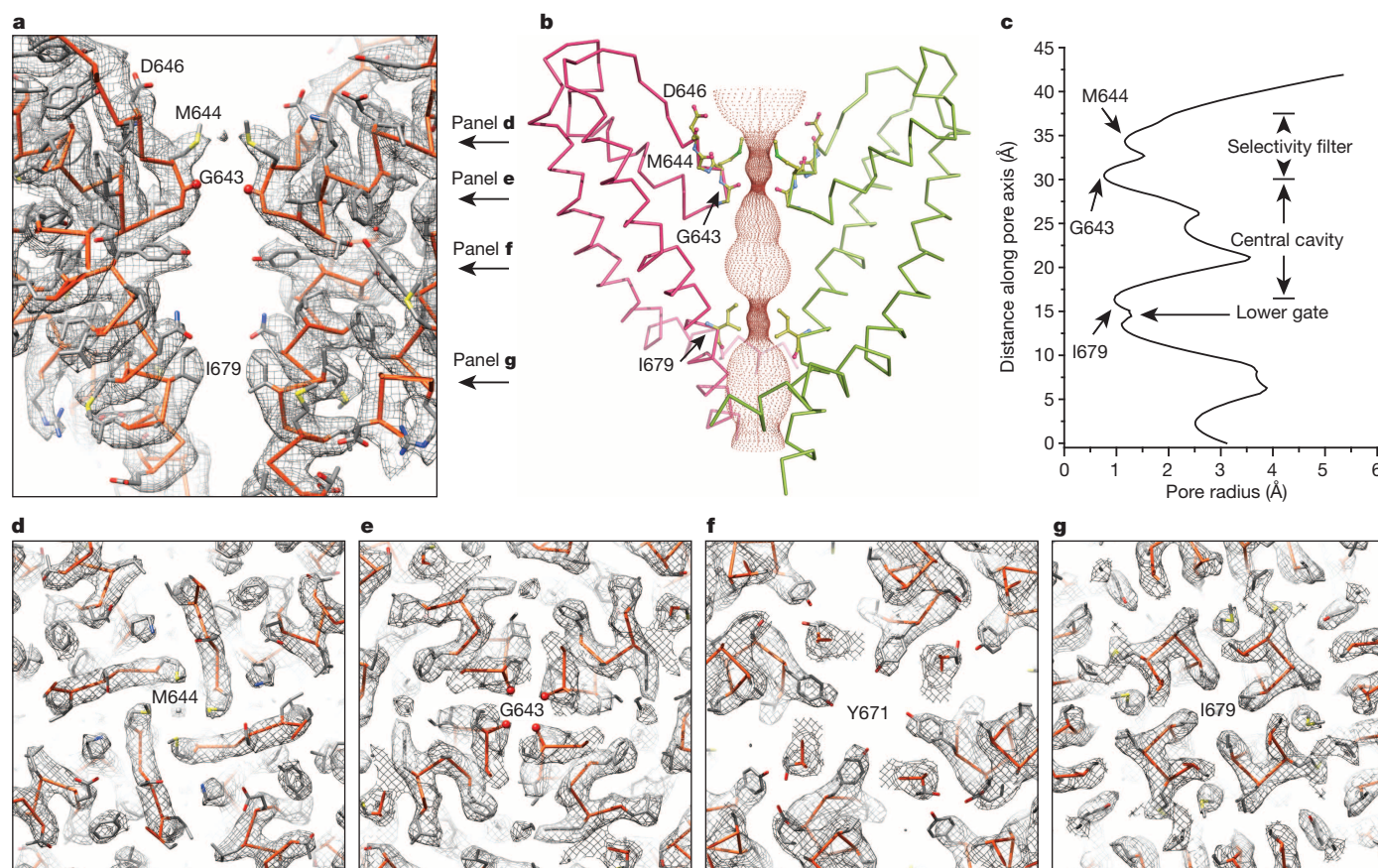


Figure 5 | The ion permeation pathway of TRPV1. **a**, Cryo-EM densities of the pore in longitudinal cross section are superimposed on an atomic model. Only two diagonally opposed subunits are shown for clarity. Several residues along the pore are labelled for orientation. Arrows denote positions of density maps for horizontal cross sections shown in panels **d–g**, as indicated. **b**, Solvent-accessible pathway along the pore mapped using the HOLE

program. Residues located at the selectivity filter and lower gate are rendered as sticks. **c**, radius of the pore calculated with program HOLE. **d–g**, Cryo-EM densities of several residues along the pore are superimposed on the atomic model; all panels represent views along the four-fold axis, showing residues from each subunit of the homotetrameric channel.

skirt, much of which remains available for interaction with other, as yet unidentified cellular factors. Whether and how these regions contribute to channel regulation may be revealed by structures of TRPV channels in complex with physiological modulators that target this domain.

Concluding remarks

TRP channels represent one of the last major ion channel families to yield to structure determination at the atomic level. Obtaining crystals

for membrane proteins, particularly of mammalian origin, is generally challenging, perhaps especially so for TRP channels, which respond to diverse stimuli (chemical and physical) and are therefore believed to be conformationally dynamic. If so, then this represents an additional obstacle to coaxing these proteins into forming well-ordered crystal lattices required for X-ray or electron crystallographic analysis. We have circumvented this problem by taking advantage of recent technological breakthroughs in single-particle cryo-EM to gain high-resolution

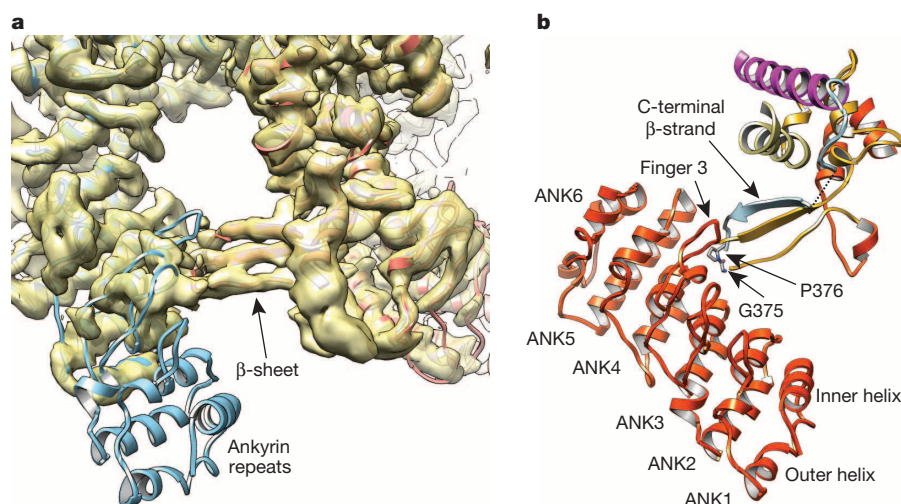


Figure 6 | Cytosolic interactions mediated by ARDs. **a**, Cryo-EM density is well fitted by crystal structure of TRPV1 ankyrin repeats 3–6 (PDB 2PNN) as a rigid body. Ankyrin repeats 1 and 2 are not observed, presumably owing to high flexibility. **b**, Finger 3 and inner helices from ankyrin repeat 3 and 4 on the concave surface of the ARD interact with β -strands from the linker and C terminus of an adjacent subunit.

structural information of purified TRP channels under conditions that do not require absolute conformational homogeneity (supported by the fact that only a subset of the data set was used to generate the final 3D reconstruction), or large amounts of material. We expect that this approach will be particularly powerful for analysing heterogeneous populations of channels with the goal of identifying and characterizing different conformational states en route to gating, or in complex with agonists, drugs or other modulators. Indeed, we provide a demonstration of this in the accompanying study²⁹.

METHODS SUMMARY

A minimal-functional rat TRPV1 construct was cloned into a modified BacMam vector (Invitrogen) containing an N-terminal fusion cassette (Kozac-MBP-tobacco etch virus (TEV) protease site) for purification on amylose resin. TRPV1 protein was expressed in HEK293S GnTI⁻ cells grown in suspension at 37 °C. Cells were collected 48 h after transduction for preparation of crude membrane and subsequent protein purification, as described¹⁰. Negative-stain EM and cryo-EM were carried out following established protocols^{15,44}. At 3.4 Å resolution, the cryo-EM map was of sufficient quality for *de novo* atomic model building. A poly-alanine model was first built in Coot and amino acid assignment subsequently achieved based mainly on the clearly defined side-chain densities of bulky residues such as Phe, Tyr and Trp, as well as some Arg and Lys residues.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 July; accepted 30 October 2013.

- Ramsey, I. S., Delling, M. & Clapham, D. E. An introduction to TRP channels. *Annu. Rev. Physiol.* **68**, 619–647 (2006).
- Venkatachalam, K. & Montell, C. TRP channels. *Annu. Rev. Biochem.* **76**, 387–417 (2007).
- Nilius, B. & Owsianik, G. Transient receptor potential channelopathies. *Pflugers Arch.* **460**, 437–450 (2010).
- Nilius, B. *et al.* Gating of TRP channels: a voltage connection? *J. Physiol. (Lond.)* **567**, 35–44 (2005).
- Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel. *Science* **309**, 897–903 (2005).
- Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. Atomic structure of a voltage-dependent K⁺ channel in a lipid membrane-like environment. *Nature* **450**, 376–382 (2007).
- Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475**, 353–358 (2011).
- Zhang, X. *et al.* Crystal structure of an orthologue of the NaChBac voltage-gated sodium channel. *Nature* **486**, 130–134 (2012).
- Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
- Cao, E., Cordero-Morales, J. F., Liu, B., Qin, F. & Julius, D. TRPV1 channels are intrinsically heat sensitive and negatively regulated by phosphoinositide lipids. *Neuron* **77**, 667–679 (2013).
- Yao, J., Liu, B. & Qin, F. Kinetic and energetic analysis of thermally activated TRPV1 channels. *Biophys. J.* **99**, 1743–1753 (2010).
- Brederson, J. D., Kym, P. R. & Szallasi, A. Targeting TRP channels for pain relief. *Eur. J. Pharmacol.* **716**, 61–76 (2013).
- Julius, D. TRP channels and pain. *Annu. Rev. Cell Dev. Biol.* **29**, 355–384 (2013).
- Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* **2**, e00461 (2013).
- Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
- Yu, X., Jin, L. & Zhou, Z. H. 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature* **453**, 415–419 (2008).
- Zhang, X., Jin, L., Fang, Q., Hui, W. H. & Zhou, Z. H. 3.3 Å cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141**, 472–482 (2010).
- Mio, K. *et al.* The TRPC3 channel has a large internal chamber surrounded by signal sensing antennas. *J. Mol. Biol.* **367**, 373–383 (2007).
- Moiseenkova-Bell, V. Y., Stanciu, L. A., Serysheva, I. I., Tobe, B. J. & Wensel, T. G. Structure of TRPV1 channel revealed by electron cryomicroscopy. *Proc. Natl Acad. Sci. USA* **105**, 7451–7455 (2008).
- Shigematsu, H., Sokabe, T., Danev, R., Tominaga, M. & Nagayama, K. A 3.5-nm structure of rat TRPV4 cation channel revealed by Zernike phase-contrast cryoelectron microscopy. *J. Biol. Chem.* **285**, 11210–11218 (2010).
- Chung, M. K., Guler, A. D. & Caterina, M. J. TRPV1 shows dynamic ionic selectivity during agonist stimulation. *Nature Neurosci.* **11**, 555–564 (2008).
- Myers, B. R., Bohlen, C. J. & Julius, D. A yeast genetic screen reveals a critical role for the pore helix domain in TRP channel gating. *Neuron* **58**, 362–373 (2008).
- Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
- Lishko, P. V., Procko, E., Jin, X., Phelps, C. B. & Gaudet, R. The ankyrin repeats of TRPV1 bind multiple ligands and modulate channel sensitivity. *Neuron* **54**, 905–918 (2007).
- Catterall, W. A. Ion channel voltage sensors: structure, function, and pathophysiology. *Neuron* **67**, 915–928 (2010).
- Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **456**, 891–897 (2008).
- Boukalova, S., Marsakova, L., Teisinger, J. & Vlachova, V. Conserved residues within the putative S4–S5 region serve distinct functions among thermosensitive vanilloid transient receptor potential (TRPV) channels. *J. Biol. Chem.* **285**, 41455–41462 (2010).
- Boukalova, S., Teisinger, J. & Vlachova, V. Protons stabilize the closed conformation of gain-of-function mutants of the TRPV1 channel. *Biochim. Biophys. Acta* **1833**, 520–528 (2013).
- Cao, E., Liao, M., Cheng, Y. & Julius, D. TRPV1 structures in distinct conformations reveal mechanisms of activation. *Nature* <http://dx.doi.org/10.1038/nature12823> (this issue).
- Latorre, R., Zaelzer, C. & Brauchi, S. Structure–functional intimacies of transient receptor potential channels. *Q. Rev. Biophys.* **42**, 201–246 (2009).
- Lin, Z. *et al.* Exome sequencing reveals mutations in TRPV3 as a cause of Olmsted syndrome. *Am. J. Hum. Genet.* **90**, 558–564 (2012).
- Long, S. B., Campbell, E. B. & MacKinnon, R. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* **309**, 903–908 (2005).
- Loukin, S., Su, Z. & Kung, C. Increased basal activity is a key determinant in the severity of human skeletal dysplasia caused by TRPV4 mutations. *PLoS ONE* **6**, e19533 (2011).
- Owsianik, G., Talavera, K., Voets, T. & Nilius, B. Permeation and selectivity of TRP channels. *Annu. Rev. Physiol.* **68**, 685–717 (2006).
- Susankova, K., Ettrich, R., Vyklicky, L., Teisinger, J. & Vlachova, V. Contribution of the putative inner-pore region to the gating of the transient receptor potential vanilloid subtype 1 channel (TRPV1). *J. Neurosci.* **27**, 7578–7585 (2007).
- Voets, T., Janssens, A., Droogmans, G. & Nilius, B. Outer pore architecture of a Ca²⁺-selective TRP channel. *J. Biol. Chem.* **279**, 15223–15230 (2004).
- Binshtok, A. M., Bean, B. P. & Woolf, C. J. Inhibition of nociceptors by TRPV1-mediated entry of impermeant sodium channel blockers. *Nature* **449**, 607–610 (2007).
- Salazar, H. *et al.* Structural determinants of gating in the TRPV1 channel. *Nature Struct. Mol. Biol.* **16**, 704–710 (2009).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Gaudet, R. A primer on ankyrin repeat function in TRP channels and beyond. *Mol. Biosyst.* **4**, 372–379 (2008).
- Inanobe, A., Matsuura, T., Nakagawa, A. & Kurachi, Y. Structural diversity in the cytoplasmic region of G protein-gated inward rectifier K⁺ channels. *Channels (Austin)* **1**, 39–45 (2007).
- Nishida, M. & MacKinnon, R. Structural basis of inward rectification: cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 Å resolution. *Cell* **111**, 957–965 (2002).
- Inada, H., Procko, E., Sotomayor, M. & Gaudet, R. Structural and biochemical consequences of disease-causing mutations in the ankyrin repeat domain of the human TRPV4 channel. *Biochemistry* **51**, 6195–6206 (2012).
- Booth, D. S., Avila-Sakar, A. & Cheng, Y. Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *J. Vis. Exp.* **58**, 3227 (2011).

Acknowledgements We thank X. Li for assistance with data acquisition using TF30 Polara and K2 Summit camera, S. Zhou and D. King for help with protein microsequencing and J.P. Armache, C. Bohlen, J. Cordero-Morales and J. Osteen for discussion and reading of the manuscript. This work was supported by grants from the National Institutes of Health (R01GM098672 and S10RR026814 to Y.C. and R01NS065071 and R01NS047723 to D.J.), the National Science Foundation (DBI-0960271 to D. Agard and Y.C.) and the University of California, San Francisco Program for Breakthrough Biomedical Research (Y.C.). E.C. was a fellow of the Damon Runyon Cancer Research Foundation.

Author Contributions All authors designed experiments. E.C. expressed and purified all protein samples used in this work and performed all functional studies. M.L. carried out all cryo-EM experiments, including data acquisition and processing. E.C. built the atomic model on the basis of cryo-EM maps. All authors analysed data and wrote the manuscript.

Author Information 3D cryo-EM density map of TRPV1 complexes without low-pass filter and amplitude modification have been deposited in the Electron Microscopy Data Bank under the accession number EMD-5778 (TRPV1). Particle images related to this entry are available for download at <http://www.ebi.ac.uk/~ardan/aspera/em-aspera-demo.html> with identification no. 10005. The coordinates of atomic model of TRPV1 have been deposited in the Protein Data Bank under the accession number 3J5P. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J. (david.julius@ucsf.edu) or Y.C. (ycheng@ucsf.edu).

METHODS

Protein expression and purification. A BacMam vector was generated from pFastBac1 (Invitrogen) to direct protein expression in mammalian cells after baculovirus transduction⁴⁵. In brief, the polyhedron promoter (P_{PH}) in pFastBac1 was replaced with a mammalian cell active promoter (P_{CMV}), immediately followed by an N-terminal fusion cassette (Kozac-MBP-tobacco etch virus (TEV) protease site) for affinity purification with amylose resin (New England Biolabs). A minimal functional rat TRPV1 construct, composed of amino acids 110 to 603 and 627 to 764, was cloned into this modified BacMam vector, and recombinant baculoviruses obtained following the manufacturer's protocol (Bac-to-Bac expression system, Invitrogen). For protein expression, HEK293S GnTI⁻ cells⁴⁶, grown in suspension at 37 °C in an orbital shaker, were transduced when cell density reached $\sim 2 \times 10^6$ per ml. Sodium butyrate was added to the culture 24 h after transduction at a final concentration of 10 mM to boost protein expression, and cells collected 48 h after transduction for preparation of crude membrane and subsequent protein purification, as described¹⁰ with slight modification. TRPV1 channel was eluted from amylose resin with buffer composed of 150 mM NaCl, 2 mM *tris*(2-carboxyethyl)-phosphine (TCEP), 10% glycerol, 20 mM HEPES, 0.5 mM DDM, 0.1 mg ml⁻¹ soybean lipids, and 20 mM maltose, then incubated with TEV protease for 4 h at 4 °C. The cleaved protein sample was then mixed with amphipols at 1:3 (w/w) with gentle agitation for another 4 h. Detergent was removed with Bio-Beads SM-2 (4 °C overnight, 15 mg per 1 ml channel/detergent/amphipols mixture). Bio-beads were then removed over a disposable polypropylene column, and eluent cleared by centrifugation before further separation on a Superdex 200 column in buffer composed of 150 mM NaCl, 20 mM HEPES, 2 mM TCEP, pH 7.4. The peak corresponding to tetrameric TRPV1 channels was collected for analysis by cryo-EM.

Cell imaging and electrophysiology. Forty-eight hours after transduction, HEK293S GnTI⁻ cells were loaded with Fura-2-acetoxymethyl ester in physiologic Ringer's buffer (140 mM NaCl, 5 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, 10 mM HEPES, pH 7.4) for ratiometric calcium imaging. Currents were recorded at room temperature in the whole-cell patch-clamp configuration. The intracellular solution contained 150 mM NaCl, 10 mM HEPES, 10 mM EGTA, pH 7.4. DkTx and capsaicin were freshly diluted into this buffer and applied via an in-line perfusion system (Automatic Scientific). To obtain proton-evoked current, buffer consisting of 150 mM NaCl, 10 mM sodium acetate, 10 mM EGTA, pH 5.0 was perfused onto TRPV1-expressing cells. Purified TRPV1 protein was also reconstituted into soybean lipid liposomes, and activation by capsaicin or heat examined by patch-clamp analysis, as described¹⁰. To estimate ion permeability ratios, the intracellular solution contained: 150 mM NaCl, 10 mM HEPES, 10 mM EDTA, pH 7.4, and extracellular solutions contained: (A) 150 mM KCl, 10 mM HEPES, 10 mM EDTA, pH 7.4; (B) 150 mM CsCl, 10 mM HEPES, 10 mM EDTA, pH 7.4; (C) 110 mM CaCl₂, 10 mM HEPES, pH 7.4. Resiniferatoxin, capsaicin and DkTx were diluted in appropriate solution immediately before use, and applied to the TRPV1-expressing cell using an in-line perfusion system.

EM data acquisition. Detergent solubilized TRPV1 particles were monodispersed as assessed by negative-stain EM (Extended Data Fig. 3a, b), enabling us to obtain a 3D reconstruction by random conical tilt (Extended Data Fig. 3c, d), which served as an initial model for cryo-EM 3D refinement. Grids for negative-stain EM were prepared following the established protocol⁴⁴. Specifically, 2.5 μ l of purified TRPV1 was applied to glow-discharged EM grids covered by a thin layer of continuous carbon film and was stained with 0.75% (w/v) uranyl formate. Negatively stained EM grids were imaged on a Tecnai T12 microscope (FEI Company) operated at 120 kV. Images were recorded at a nominal magnification of 67,000 \times using a 4k \times 4k charge-coupled device camera (UltraScan 4000, Gatan), corresponding to a pixel size of 1.73 Å per pixel on the specimen. Tilt pair images for random conical tilt 3D reconstruction were manually recorded at 50° and 0°.

For cryo-EM, detergent was replaced by amphipols^{47,48}, in which purified TRPV1 remained stable and monodispersed (Extended Data Fig. 3e–g). 2 μ l of purified TRPV1 sample at a concentration of ~ 0.3 mg ml⁻¹ was applied to a glow-discharged Quantifoil holey carbon grid (1.2 μ m hole size, 400 mesh), blotted inside a Vitrobot Mark III (FEI Company) using 6-s blotting time with 90% humidity, and then plunge-frozen in liquid ethane cooled by liquid nitrogen. Cryo-EM images were collected at liquid nitrogen temperature on a Tecnai TF20 electron microscope (FEI) operated at 200 kV using a CT3500 side entry holder (Gatan), following the low-dose procedure. On the TF20 microscope, images were recorded at a nominal magnification of 80,000 \times using a phosphor scintillator based TemF816 8K \times 8K CMOS camera (TVIPS GmbH), corresponding to a pixel size of 0.9 Å per pixel on the specimen. Images were recorded with a defocus in the range from 1.5 to 3.5 μ m (Extended Data Fig. 4). We determined a 3D reconstruction with C4 symmetry to an overall resolution of 8.8 Å, using gold-standard FSC = 0.143 criteria²³ (Extended Data Fig. 5a). In this density map, transmembrane α -helices are clearly resolved. The shape of the cytoplasmic domain is clearly

defined, but secondary structural features are not well resolved (Extended Data Figs 5b–f), probably indicating greater flexibility of this region.

Another data set of frozen hydrated TRPV1 particles were collected on a TF30 Polara electron microscope (FEI Company) operated at 300 kV. It is equipped with a K2 Summit direct electron detector camera (Gatan). Images were recorded using super-resolution counting mode following an established protocol¹⁵. Specifically, images from TF30 were recorded at a nominal magnification of 31,000 \times , corresponding to a calibrated super resolution pixel size of 0.6 Å per pixel on the specimen. The dose rate on the camera was set to be ~ 8 counts (corresponding to ~ 9.9 electron) per physical pixel per second. The total exposure time was 6 s, leading to a total accumulated dose of 41 electrons per Å² on the specimen. Each image was fractionated into 30 subframes, each with an accumulation time of 0.2 s per frame. All dose-fractionated cryo-EM images were recorded using a semi-automated acquisition program UCSFImage4 (written by X. Li). Images were recorded with a defocus in a range from 1.5 to 3.0 μ m.

Image processing. SamViewer, an interactive image analysis program written in wxpython, was used for all 2D image display and particle picking. Negative-stain EM images were 2 \times 2 binned for manual particle picking. Defocus was determined using CTFFIND and CTFILT⁴⁹. Individual particles were cut out and normalized to have a mean of 0 and a standard deviation of 1. For 2D classification, particles were first corrected for contrast transfer function (CTF) by flipping the phase using 'ctffaply' (written by X. Li), and subjected to 10 cycles of correspondence analysis, *k*-means classification and multi-reference alignment (MRA), using SPIDER operations 'CA S', 'CL KM' and 'AP SH'⁵⁰. For random conical tilt (RCT) 3D reconstruction, SamViewer was used for picking particles from the tilt-pair images, as well as determination of the tilting axes and angles. After 2D classification of untitled particles, RCT 3D reconstructions of each 2D class were calculated using FREALIGN⁵¹.

Low-dose images of frozen hydrated TRPV1 collect on TF20 were binned 2 \times 2, resulting with a pixel size of 1.9 Å, for image processing. For particle picking and 2D classification, images were 2 \times 2 binned further to a pixel size of 3.8 Å. Dose-fractionated super-resolution image stacks of frozen hydrated TRPV1 images collected using K2 Summit camera were first binned 2 \times 2 resulting with a pixel size of 1.2 Å for motion correction and further image processing. After motion correction¹⁵, a sum of all subframes in each image stack was used for further processing. Particle picking and 2D classification used 6 \times 6 binned images (3.6 Å per pixel), and 3D classification used 4 \times 4 binned images (2.4 Å per pixel). Final 3D reconstruction was calculated from 2 \times 2 binned images (1.2 Å per pixel). Image binning was calculated using Fourier cropping. First, $\sim 2,000$ particles were picked interactively and classified into ~ 10 2D classes using the same classification procedure described above. Then, for each micrograph, the entire image was cut into a set of overlapping small images with a window of 64 \times 64 pixels, all windowed images were subject to MRA against the 2D class averages generated from manually selected particles (implemented in a python script, 'samautopick.py'). All particles were then displayed in SamViewer, in the order of their cross-correlation values from MRA, and a threshold was interactively set to remove the particles with cross-correlation values below the threshold. A small number of particles with blurry appearance or with obvious wrong shape and size were also removed interactively. We picked 70,585 particles from 300 cryo-EM images collected on the TF20, and 97,166 particles from 946 image stacks collected on the TF30. The selected particles were further screened by a reference-free 2D classification. A total of 45,625 particles from the TF20 data and 88,915 particles from the TF30 data were kept for determinations of 3D reconstructions.

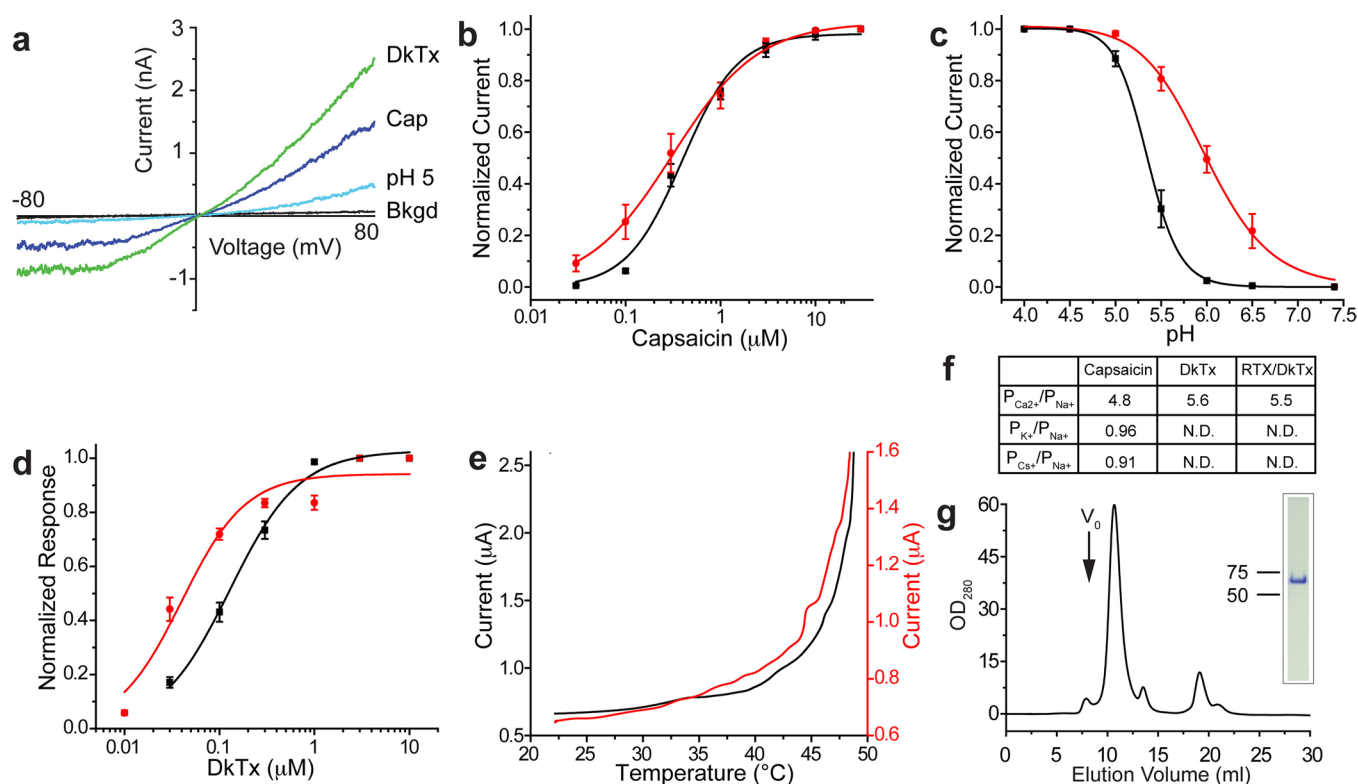
All 3D classification and refinement were carried out in RELION⁵². The RCT 3D reconstruction from negatively stained TRPV1 was low-pass filtered to 60 Å, and used as the starting model for 3D classification of the TF20 data. After 25 iterations of 3D classification with 6 classes, particles in the classes that produced similar 3D reconstructions were combined for 3D auto-refinement. In this step, the 3D reconstructions from the 3D classification were low-pass filtered to 60 Å and used as the starting models. 10,357 particles were used to calculate the final map (Extended Data Fig. 5). The resolution was estimated by the gold-standard FSC = 0.143 criteria, after applying a soft spherical mask on the two reconstructions refined from the half-data sets independently²³. The resolution of the 3D reconstruction from TF20 data was estimated as 8.8 Å. This 3D reconstruction was again low-pass filtered to 60 Å, and used as the starting model for 3D classification of the TF30 Polara data. 3D classification and refinement were carried out using the same procedure as for the TF20 data. The 3D reconstruction calculated from selected 35,645 particles each averaged from 30 subframes of motion-corrected image stack reached a resolution of 3.6 Å. Continued refinement using the same particles but averaged from subframe 3–16 improved the resolution further to 3.4 Å. This was because the first 2 frames contain the most severe motion and the later frames accumulate more radiation damage¹⁵. The accumulated dose of the first 16 frames is 21 e⁻/Å² on specimen. Final gold-standard FSC curve was calculated

using a soft spherical mask (with a 5-pixel fall-off) on the two independent reconstructions (Extended Data Fig. 8). The accuracies of rotation and translation reported by RELION were 3.54° and 1.358 Å. RELION post-processing with auto-mask and auto-bfactor determined the resolution of the final map as 3.28 Å and bfactor as -101.228 Å^2 . For model building and visualization, amplitude of the final 3D density map of 3.4 Å resolution was amplified either by a temperature factor, -100 Å^2 (Fig. 5) or by frequency-dependent scaling factor determined by comparing the experimental 3D density map with the ideal density map calculated from the atomic model (Figs 1d–f, 6a)¹⁵. All maps shown in the figures were low-pass filtered to the stated resolutions. The map deposited to the EMDB database is the raw map without amplitude sharpening, masking or filtering. UCSF Chimera⁵³ was used to visualize and segment the cryo-EM maps, and the rigid body fitting of atomic model was done using the fit-in-map function from Chimera.

Model building in Coot. At 3.4 Å resolution, the cryo-EM map was of sufficient quality for *de novo* atomic model building, except for the ARD, where the density is not as well resolved as other regions of the channel. Taking advantage of the defined geometry of α -helices and clear bumps for C α atoms, a polyaniline model was first built in Coot^{54,55}. Amino acid assignment was subsequently achieved based mainly on the clearly defined side chains densities of bulky residues such as Phe, Tyr and Trp, as well as some Arg and Lys residues. The ARD structure (PDB 2PNN) was docked into the density as a rigid body without further modification. The final model includes almost all residues, except for those located at S2–S3 loop (503–507) and a region C-terminal to the TRP domain (720–751). Although a C-terminal region of the channel (registered as 752–762 in the model) clearly assumes a β -strand secondary structure, amino acids could not be assigned since side chain density is not well resolved at this segment. The final model exhibits good geometry as indicated by the Ramachandran plot (preferred region, 95.73%; allowed region, 3.92%; outliers, 0.34%). Of note, the six transmembrane helices, TRP domain and pre-S1 helix are devoid of any such outliers. One outlier is located

at the S3–S4 loop and the other residues at docked ankyrin repeats. Pore radii were calculated using the HOLE program⁵⁶.

45. Dukkkipati, A., Park, H. H., Waghay, D., Fischer, S. & Garcia, K. C. BacMam system for high-level expression of recombinant soluble and membrane glycoproteins for structural studies. *Protein Expr. Purif.* **62**, 160–170 (2008).
46. Reeves, P. J., Callewaert, N., Contreras, R. & Khorana, H. G. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible *N*-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc. Natl Acad. Sci. USA* **99**, 13419–13424 (2002).
47. Althoff, T., Mills, D. J., Popot, J. L. & Kuhlbrandt, W. Arrangement of electron transport chain components in bovine mitochondrial supercomplex I₁III₂IV₁. *EMBO J.* **30**, 4652–4664 (2011).
48. Popot, J. L. et al. Amphipols from A to Z. *Annu. Rev. Biophys.* **40**, 379–408 (2011).
49. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
50. Frank, J. et al. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
51. Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *J. Struct. Biol.* **157**, 117–125 (2007).
52. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
53. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
54. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
55. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
56. Smart, O. S., Neduvilil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360 (1996).



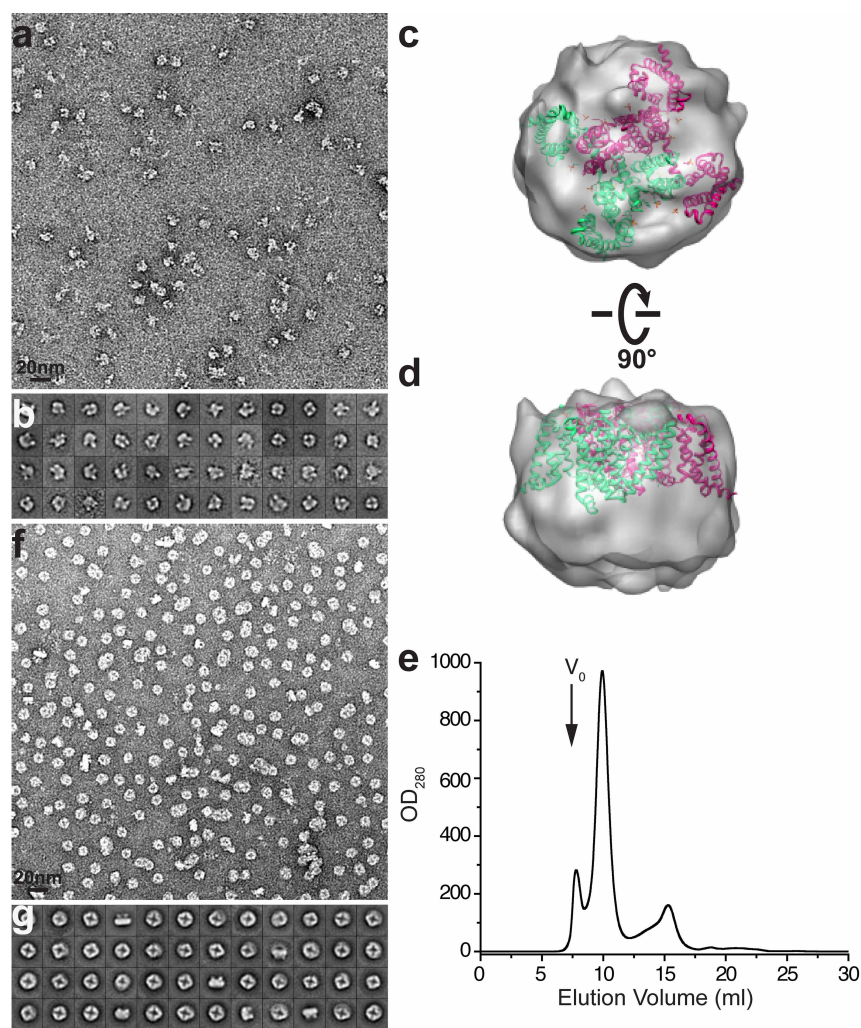
Extended Data Figure 1 | A minimal TRPV1 channel that is functional and biochemically stable. **a**, Mammalian (HEK293) cells expressing a minimal construct (with an N-terminal green fluorescent protein (GFP) tag) responded to various TRPV1 agonists, including capsaicin (Cap; $0.5 \mu\text{M}$), extracellular protons (pH 5.0) and double-knot spider toxin (DkTx; $2 \mu\text{M}$). Electrophysiological responses were measured in whole-cell patch-clamp configuration. **b**, **c**, Dose-responsive curves for capsaicin (**b**) or protons (**c**) were determined for minimal (black) or full-length (red) TRPV1, both of which contained an N-terminal GFP fusion. Values were normalized to maximal currents evoked by $30 \mu\text{M}$ capsaicin (**b**) or pH 4.0 (**c**) ($n = 6$ independent whole-cell recordings). **d**, DkTx dose-response curves for minimal (black) or full-length (red) TRPV1 as in **b** and **c**, determined by calcium imaging. Values were normalized to maximal capsaicin

($10 \mu\text{M}$)-evoked response in transfected HEK293 cells ($n > 30$ per point). **e**, Thermal response profiles for minimal (black) or full-length (red) TRPV1-expressing oocytes reveal similar heat sensitivity. **f**, Ion permeability ratios of agonist-evoked currents from minimal TRPV1 were estimated from reversal potential shifts in whole-cell patch-clamp recordings of transfected HEK293 cells, revealing no significant differences from full-length channel. **g**, Gel-filtration profile (Superdex-200) of detergent solubilized TRPV1 after purification on amylose affinity resin and proteolytic removal of maltose-binding protein (MBP) tag. The major species elutes as a symmetrical peak after the void volume (V_0). Inset shows that peak material migrates as a single, homogeneous band on SDS-PAGE (4–12% gradient gel; Coomassie stain).



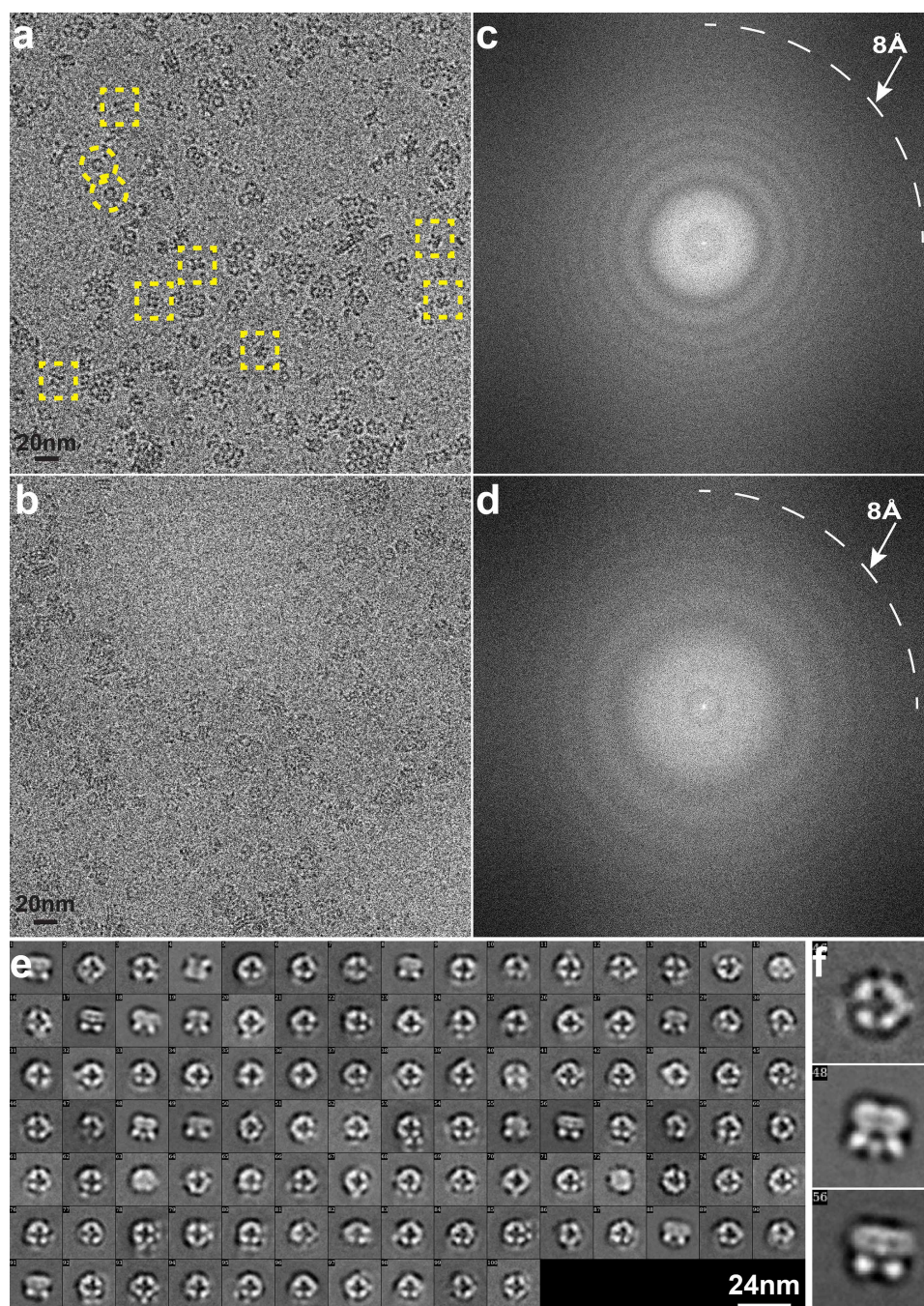
Extended Data Figure 2 | Sequence alignment of TRPV1 to other TRPV family members. The rat TRPV1 construct used for this study consists of residues 110 to 764 (indicated by red arrows), excluding the highly divergent region (604–626, highlighted by cyan box). Secondary structure elements are indicated above the sequence. The starting points of six ankyrin repeats are

based on a crystal structure of ARD of TRPV1 (PDB 2PNN). Several critical residues discussed in the text are labelled in blue, and conserved glycine and proline residues at the turn of a β -sheet (highlighted in Fig. 6) are indicated with red stars.



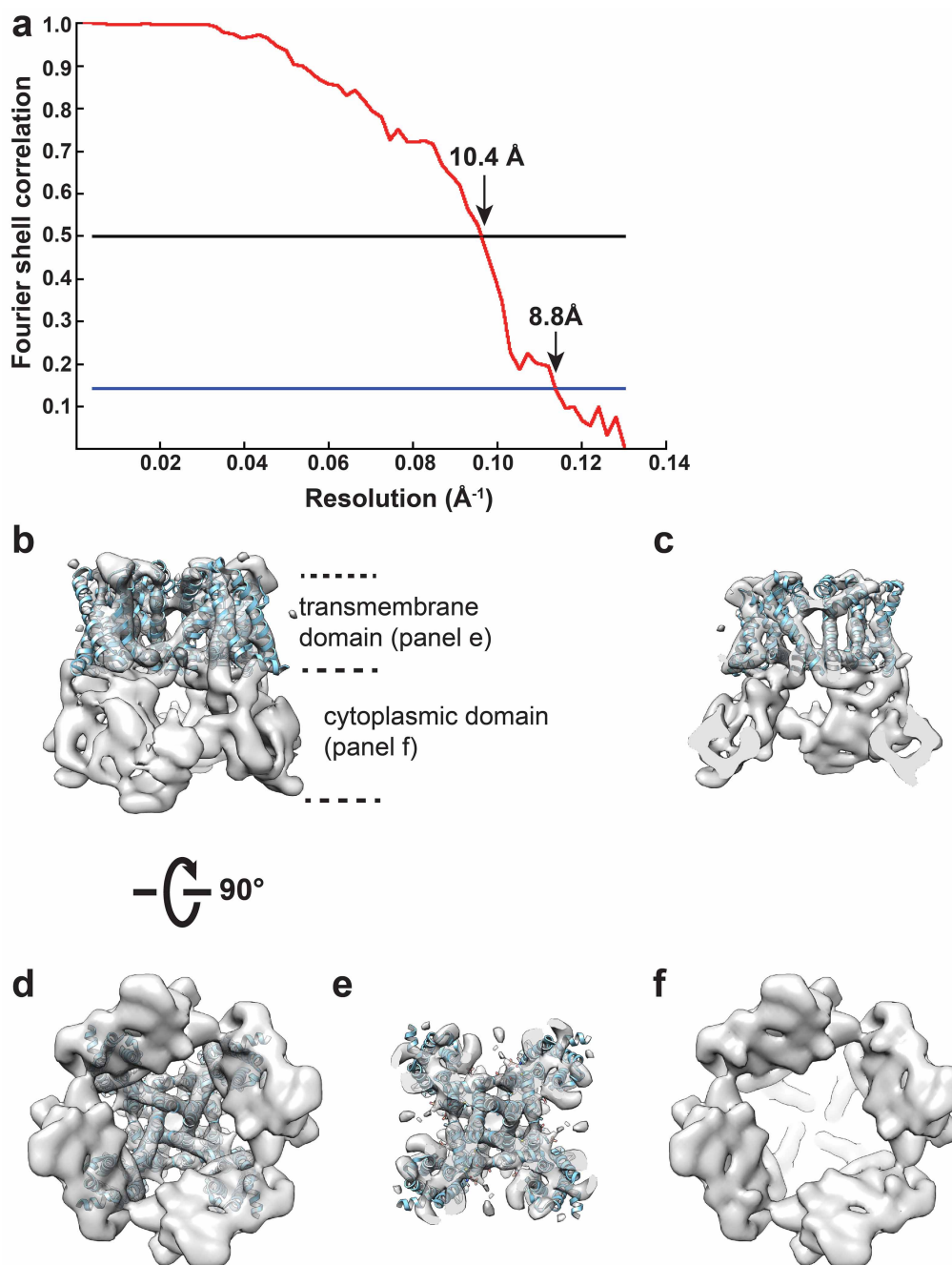
Extended Data Figure 3 | Negative-stain EM of TRPV1. **a**, Representative negative-stain image of purified minimal TRPV1 protein in detergent (n-dodecyl β -d-maltopyranoside; DDM) after proteolytic removal of MBP tag. **b**, 2D class averages of negatively stained particles in DDM. **c**, **d**, Two views of a random conical tilt (RCT) reconstruction from negatively stained TRPV1 in DDM. The RCT reconstruction was low-pass filtered at 30 Å, and fitted with

the structure of Na_vAb (PDB 3RVY) to indicate the size and general shape. **e**, Gel-filtration profile (Superdex-200) of purified minimal TRPV1 protein after exchange from DDM into amphipols. The major species elutes as a symmetrical peak after the void volume (V_0). **f**, Representative negative-stain image of purified minimal TRPV1 protein without MBP tag in amphipols. **g**, 2D class averages of negative-stain particles in amphipols.



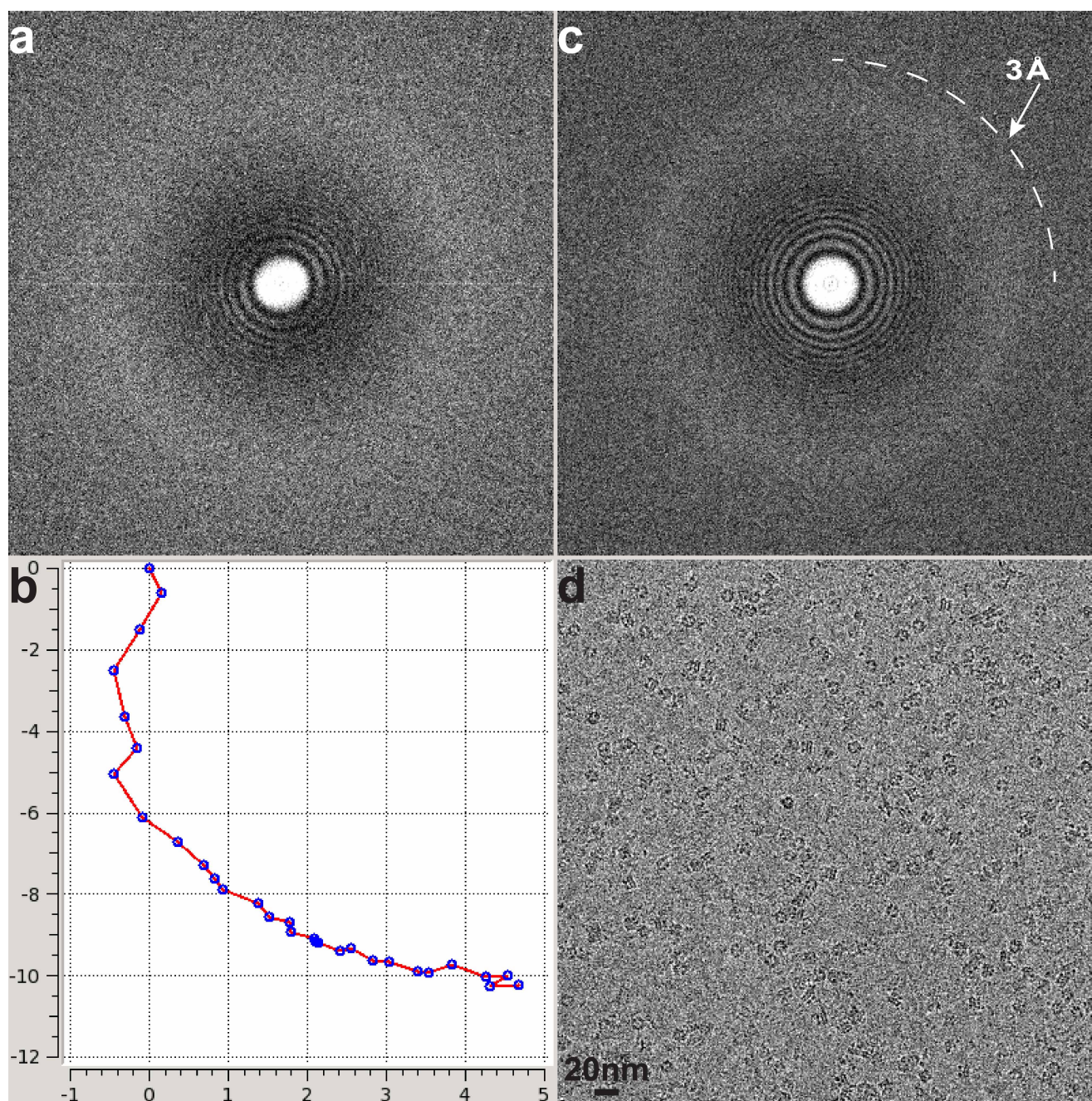
Extended Data Figure 4 | Cryo-EM of TRPV1 using Tecnai TF20 microscope and TemF816 8k × 8k CMOS camera. a–d, Representative images of frozen hydrated TRPV1 in amphipols taken at different defocus levels, 3.1 μm (a) and 1.5 μm (b) and their Fourier transforms (c, d). Thon rings

extend to ~8 Å. Dash-line squares or circles indicate representative particles showing two distinctive views. e, 2D class averages of TRPV1 particles. f, Enlarged view of three representative 2D class averages.



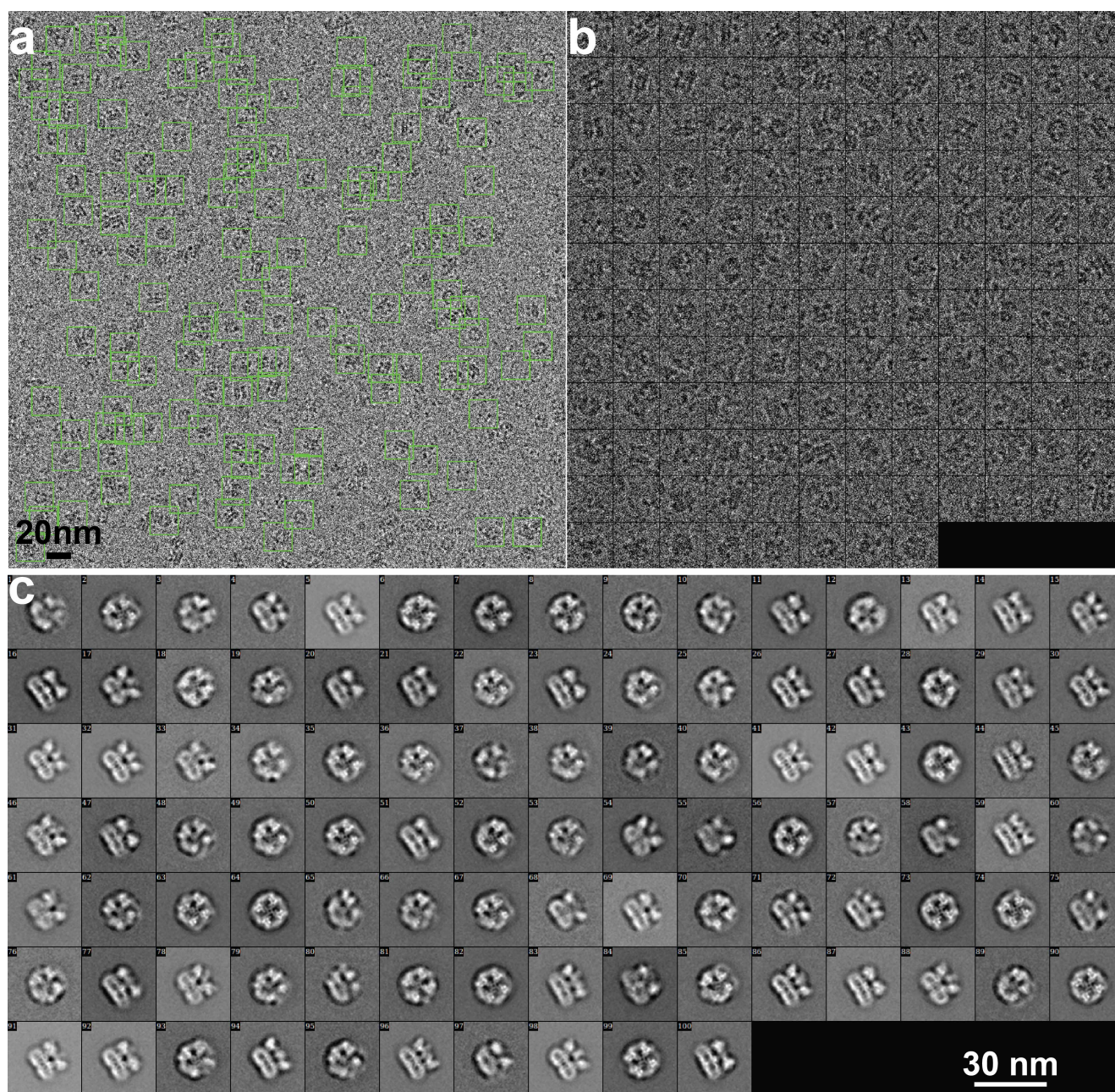
Extended Data Figure 5 | 3D reconstruction of TRPV1 calculated from TF20 data. **a**, Gold-standard FSC curve for the 3D reconstruction, marked with resolutions corresponding to $FSC = 0.5$ and 0.143 . **b**, Side view of the 3D reconstruction low-pass filtered at 9 \AA and amplified by a temperature factor $-1,500 \text{ \AA}^2$, showing transmembrane (top) and cytoplasmic (bottom) domains.

The transmembrane domain roughly fitted by the atomic model of Na_vAb (PDB 3RVY). **c**, Longitudinal cross section view focused on central transmembrane helices. **d**, Bottom-up view of the 3D reconstruction shows overall structure. **e**, **f**, Bottom-up cross-section views showing the arrangement of transmembrane (**e**) and cytoplasmic (**f**) domains.



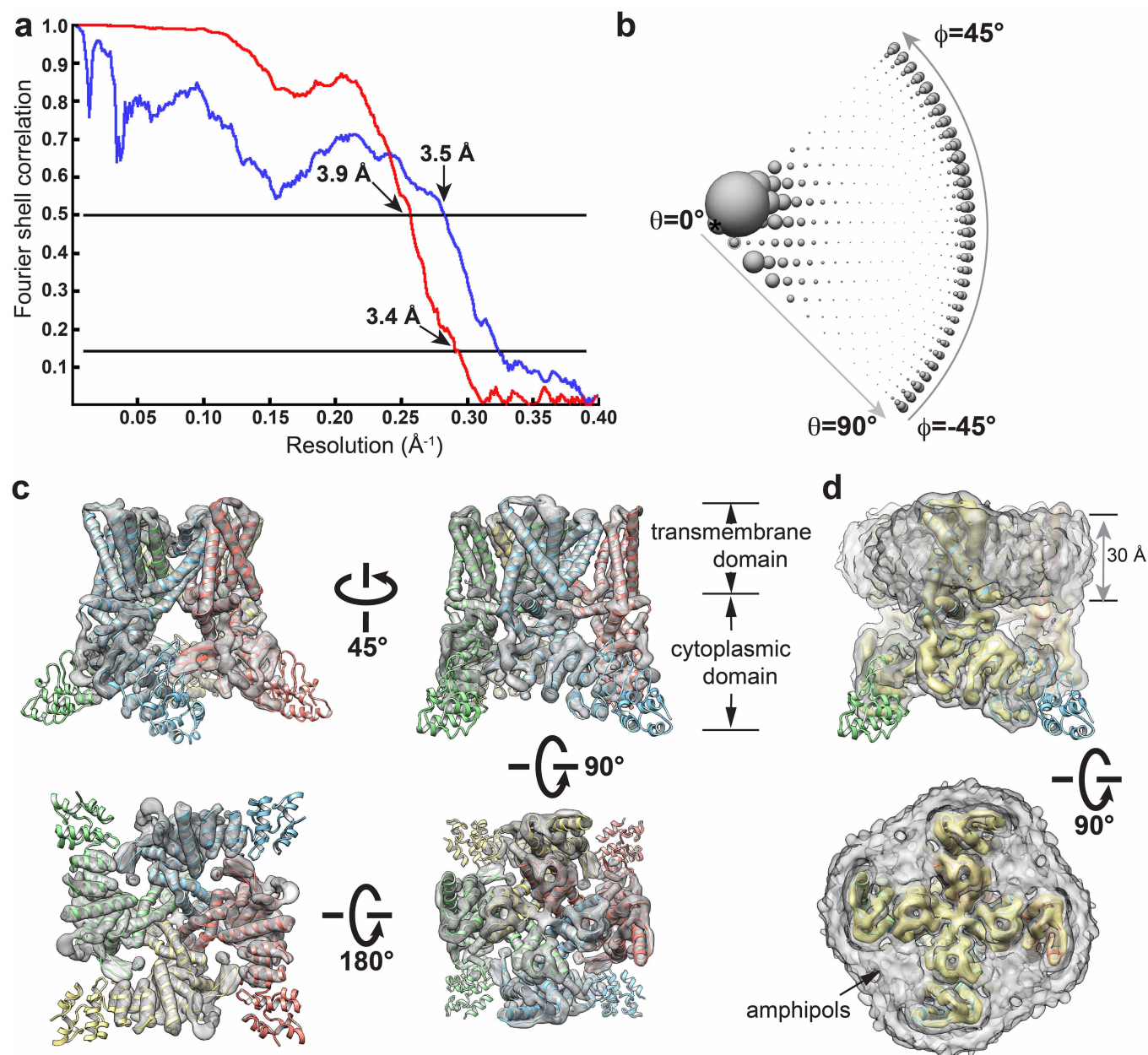
Extended Data Figure 6 | Motion correction improves the quality of images collected on Polara TF30 microscope using a K2 Summit direct electron detector. **a**, Fourier transform of a representative cryo-EM image of TRPV1 embedded in a thin layer of vitreous ice over Quantifoil hole without supporting

carbon film before motion correction. **b**, Path of motion of 30 individual subframes, determined as described in Methods. **c**, **d**, A nearly perfect Fourier transform (**c**) was restored after the EM image was corrected for motion (**d**).



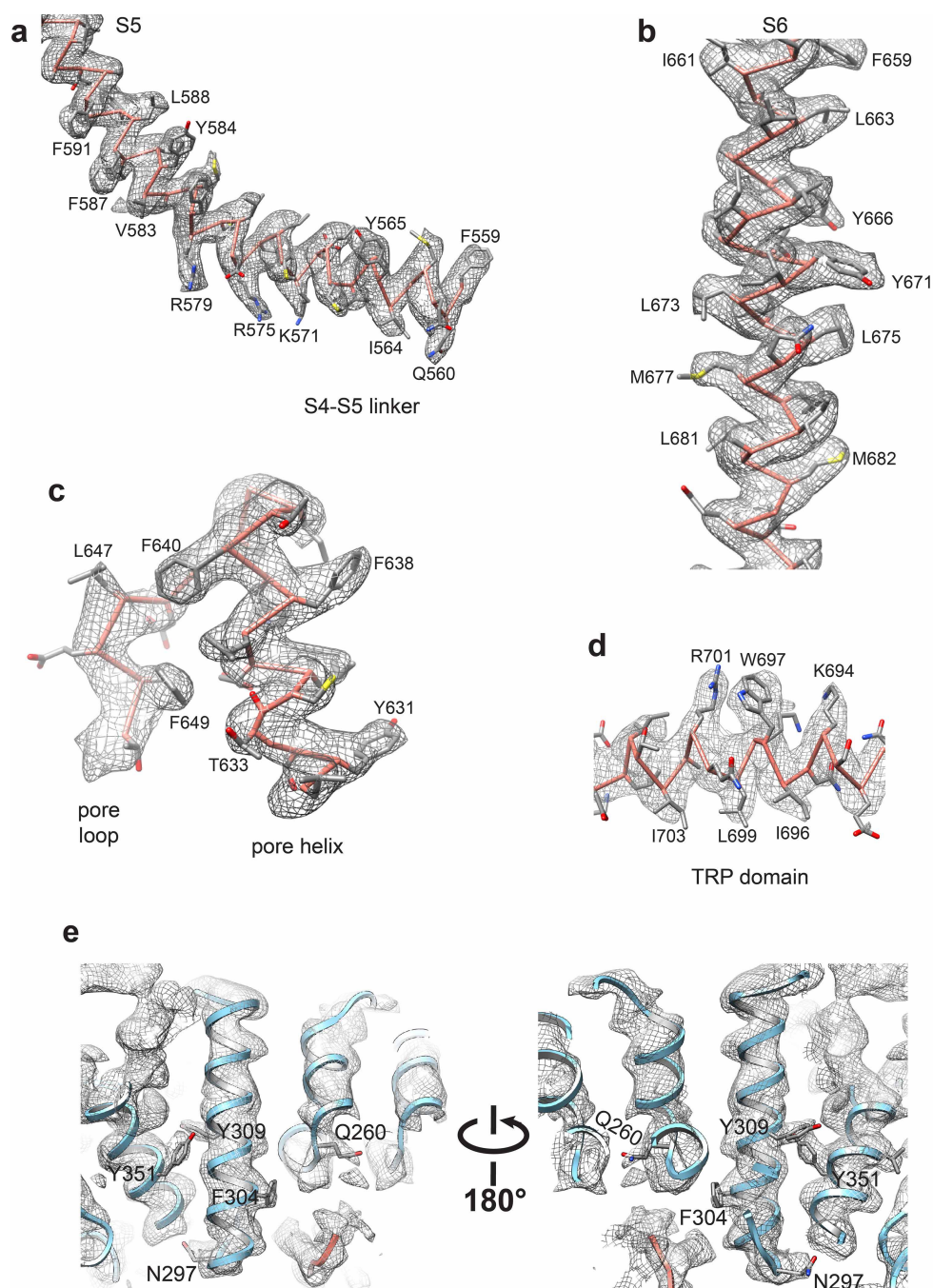
Extended Data Figure 7 | Picking and 2D classification of TRPV1 Cryo-EM particles collected on Polara TF30 microscope. **a**, Representative cryo-EM image after motion correction. Green boxes indicate all particles that were selected by semi-automatic particle picking and 2D screening, as described in

Methods. **b**, Gallery view of the particles shown in **a**. **c**, 2D class averages of cryo-EM particles show many fine features (also seen in enlarged views in Fig. 1c), and these features are not visible in the 2D class averages of cryo-EM particles from TF20 data (Extended Data Fig. 4e, f).



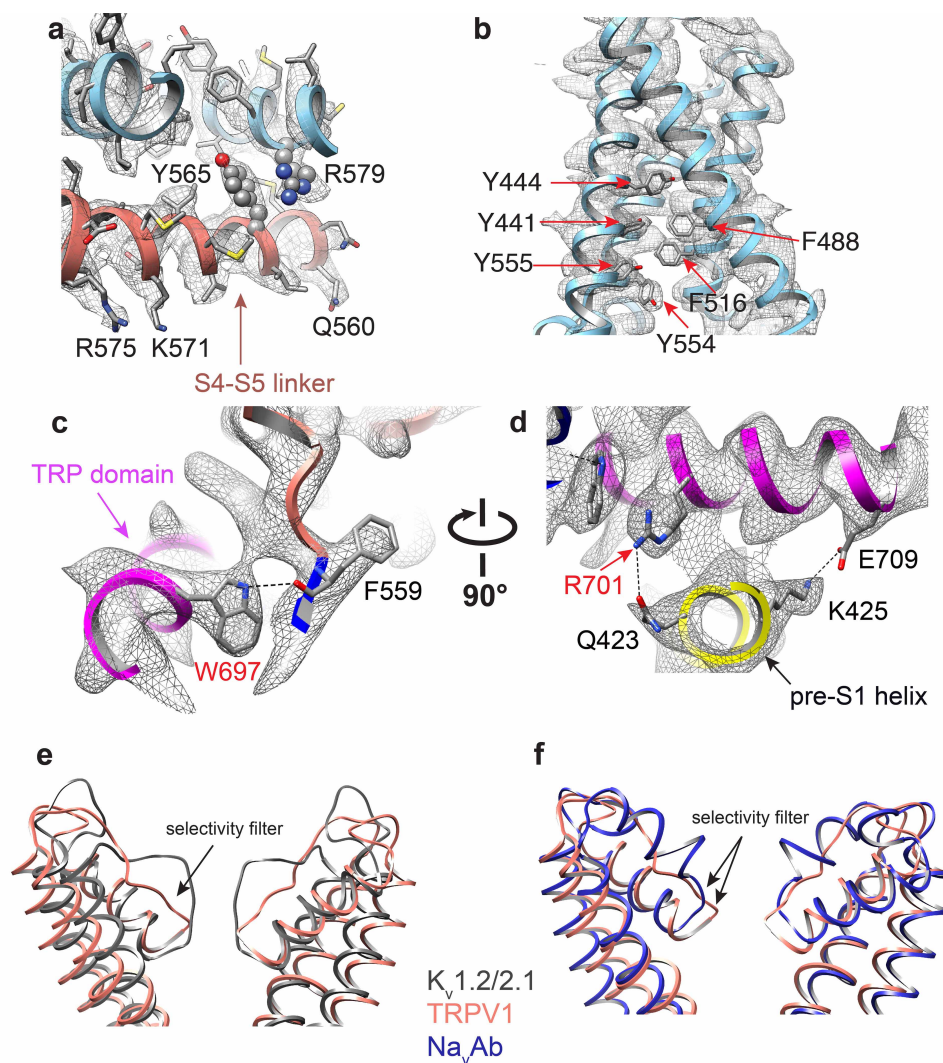
Extended Data Figure 8 | 3D reconstruction of TRPV1 calculated from TF30 data. **a**, Gold-standard FSC curve (red) of the final 3D reconstruction, marked with resolutions corresponding to $\text{FSC} = 0.5$ and 0.143 . The FSC curve between the final map and that calculated from the atomic model is shown in blue. The relative low value of this FSC (blue) at low frequency range ($>10 \text{\AA}$) is probably due to the presence of amphipol density in the experimental map. **b**, Euler angle distribution of all particles used for calculating the final 3D

reconstruction. The sizes of balls represent the number of particles. The accuracy of rotation is 3.54° , as reported by RELION. **c**, Different views of the 3D reconstruction low-pass filtered at 6 \AA and amplified by a temperature factor of -100\AA^2 , fitted with the atomic model of TRPV1. **d**, Two views of the 3D reconstruction displayed at two different isosurface levels (high in yellow and low in grey). At the low isosurface level, the belt-shaped density of amphipols is visible with a thickness of $\sim 30 \text{\AA}$.



Extended Data Figure 9 | Cryo-EM densities of selected regions of TRPV1 at 3.4 Å resolution. **a–d**, Representative cryo-EM densities (grey mesh) are superimposed on atomic model (main chain in pink) for various TRPV1 domains, as indicated. **e, f**, Representative cryo-EM densities (grey mesh) are

docked with crystal structure of TRPV1 ankyrin repeats (PDB 2PNN). Accuracy of docking was supported by fitting of several bulky side chains. Map was low-pass filtered to 3.4 Å and amplified by a temperature factor -100 Å^2 .



Extended Data Figure 10 | Details of domain interactions and outer pore configurations. **a–d**, Cryo-EM densities (grey mesh) of highlighted regions of TRPV1, as indicated, at 3.4 Å resolution are superimposed onto atomic model. Map was low-pass filtered to 3.4 Å and amplified by a temperature factor -100 Å^2 . **e**, Superimposition of TRPV1 (salmon) with Kv 1.2–2.1 chimera

(PDB 2R9R; grey). **f**, Superimposition of TRPV1 (salmon) with Na_vAb (PDB 3RVY; blue). In each case, substantial structural differences are observed in the outer pore region. Structural alignments are based on the pore domain (S5–P–S6).

TRPV1 structures in distinct conformations reveal activation mechanisms

Erhu Cao^{1*}, Maofu Liao^{2*}, Yifan Cheng² & David Julius¹

Transient receptor potential (TRP) channels are polymodal signal detectors that respond to a wide range of physical and chemical stimuli. Elucidating how these channels integrate and convert physiological signals into channel opening is essential to understanding how they regulate cell excitability under normal and pathophysiological conditions. Here we exploit pharmacological probes (a peptide toxin and small vanilloid agonists) to determine structures of two activated states of the capsaicin receptor, TRPV1. A domain (consisting of transmembrane segments 1–4) that moves during activation of voltage-gated channels remains stationary in TRPV1, highlighting differences in gating mechanisms for these structurally related channel superfamilies. TRPV1 opening is associated with major structural rearrangements in the outer pore, including the pore helix and selectivity filter, as well as pronounced dilation of a hydrophobic constriction at the lower gate, suggesting a dual gating mechanism. Allosteric coupling between upper and lower gates may account for rich physiological modulation exhibited by TRPV1 and other TRP channels.

The capsaicin (vanilloid) receptor, TRPV1, is a heat-activated cation channel that is modulated by inflammatory agents and contributes to acute and persistent pain¹. To understand the structural basis whereby TRPV1 responds to disparate physiological stimuli, it is necessary to view the channel in distinct functional states. This is an exceedingly challenging goal for eukaryotic membrane channels, having been achieved in only a handful of instances^{2–5}. Such attempts are hampered by lack of pharmacological tools with which to capture channels in specific states, and difficulties in achieving conformational uniformity required for X-ray crystallography. Among TRP channels, TRPV1 enjoys the richest pharmacology, including small molecule agonists and antagonists, as well as larger peptide toxins^{6,7}. Moreover, as described in the accompanying study, we have used single-particle electron cryo-microscopy (cryo-EM) to ascertain the structure of TRPV1 in its unliganded (apo), closed state without crystallization⁸. Together, these advantages enhance possibilities for examining this channel in multiple functional states.

TRPV1 channels are homotetramers whose three-dimensional (3D) structure resembles that of voltage-gated ion channels (VGICs), wherein an ion permeation pathway is formed by transmembrane segments 5 and 6 (S5 and S6) and the intervening pore loop region (S5–P–S6)⁸. This central pore is surrounded by four independently folded S1–S4 domains, which in the case of VGICs contain voltage sensors and undergo substantial movement during gating^{9–11}. Despite these architectural similarities, it remains unknown whether TRPs and VGICs are activated through common conformational rearrangements. One line of evidence to suggest differential gating mechanisms comes from analysis of toxins that function as gating modifiers for these channels. Spiders produce a multitude of peptide toxins¹², including some that antagonize voltage-gated potassium (K_V) channels by binding to and impeding movement of the S3–S4 voltage sensor^{13,14}. Others activate TRPV1 to elicit pain as part of the spider's chemical defence mechanism^{15,16}. One such vanillotoxin (called double-knot toxin, DkTx) is a 75-amino-acid-long peptide consisting of two independently folded toxin moieties connected by a short linker, a bivalent arrangement that enables DkTx to trap TRPV1 in its open state with near-irreversible kinetics¹⁵. Mutational

analysis suggests that DkTx binds to residues within the S5–P–S6 pore region¹⁵, consistent with the notion that the outer pore of TRPV1 is conformationally dynamic and contributes directly to gating^{15,17–19}. By contrast, the analogous region of K_V channels is believed to remain relatively stationary during normal gating^{10,20}, providing a compelling explanation for why some gating modifier toxins evolved to target the outer pore versus voltage-sensor domains of TRP and K_V channels, respectively. To resolve questions pertaining to TRP gating mechanisms and pharmacology, we determined structures of TRPV1 trapped in different conformational states by DkTx and/or small vanilloid ligands. These studies reveal differential gating mechanisms for TRPs and VGICs, while highlighting advantages of cryo-EM for capturing protein structures in distinct conformations.

Structures of TRPV1–ligand complexes

To analyse TRPV1 in activated state(s), we incubated purified ‘minimal’ channel protein⁸ with the vanilloid agonists resiniferatoxin (RTX) or capsaicin. For samples containing RTX, we also included DkTx to facilitate trapping of channels in their fully open state. We determined 3D reconstructions of TRPV1 with RTX/DkTx or capsaicin to resolutions of 3.8 Å or 4.2 Å, respectively, using gold-standard Fourier shell correlation (FSC) = 0.143 criteria²¹ (Extended Data Figs 1–4). Many side-chain densities, particularly those along the pore, were clearly resolved, allowing for accurate model building based on the structure of apo TRPV1 (Extended Data Figs 5 and 6)⁸.

In the RTX/DkTx structure, density corresponding to two DkTx molecules (that is, four toxin moieties) was observed atop each tetrameric channel complex (Fig. 1 and Extended Data Fig. 2). Mutations within extracellular loops of the TRPV1 outer pore region specifically abrogate DkTx-evoked responses¹⁵, outlining a putative vanillotoxin site. Indeed, we see that DkTx is bound to the channel in precisely this region, but in an interesting and unexpected way: each toxin moiety sits at subunit interfaces by contacting residues at the top of the pore helix from one subunit and the outer pore loop proximal to S6 from the neighbouring subunit (Figs 1 and 2a). Thus, two bivalent toxin

¹Department of Physiology, University of California, San Francisco, California 94158-2517, USA. ²Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California, San Francisco, California 94158-2517, USA.

*These authors contributed equally to this work.

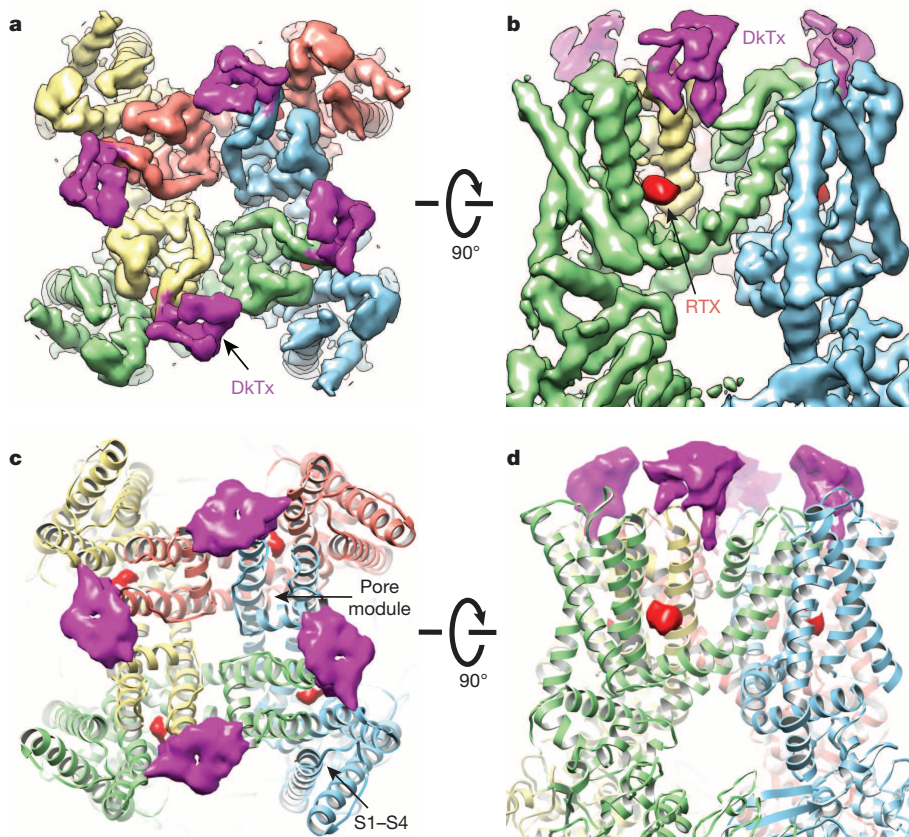


Figure 1 | Structure of TRPV1 in complex with vanilloid ligand and spider toxin. **a, b,** Cryo-EM density map of TRPV1 in complex with vanilloid agonist (resiniferatoxin, RTX; red sphere) and vanillotoxin (spider toxin, DkTx; magenta and denoted by arrow) shown as top-down (**a**) and side (**b**) views. The map is filtered to 3.8 Å and amplified with a temperature factor of -100 Å^2 . **c, d,** Structures from **a** and **b** with TRPV1 channel rendered as ribbon diagram. Density corresponding to each ligand represents signal from difference map (4.7σ) generated by comparison with apo TRPV1 structure.

molecules 'staple' the channel together in its activated state, reminiscent of interactions between psalmotoxin and homotrimeric ASIC1a channels, wherein the toxin bridges distinct regions on extracellular domains of adjacent subunits².

Mutagenesis studies have pinpointed residues that specify sensitivity to small vanilloid ligands, including Y511 and S512 in S3 and M547

and T550 in S4 (refs 22–25). Indeed, in the RTX/DkTx structure, we see robust density in close proximity to these residues, probably representing a molecule of RTX coordinated in a binding pocket just above E570 in the S4–S5 linker from the same subunit, and proximal to L669 in S6 of a neighbouring subunit (Figs 1 and 2b and Extended Data Fig. 7a). In VGICs, the S4–S5 linker has a critical role in translating movement of

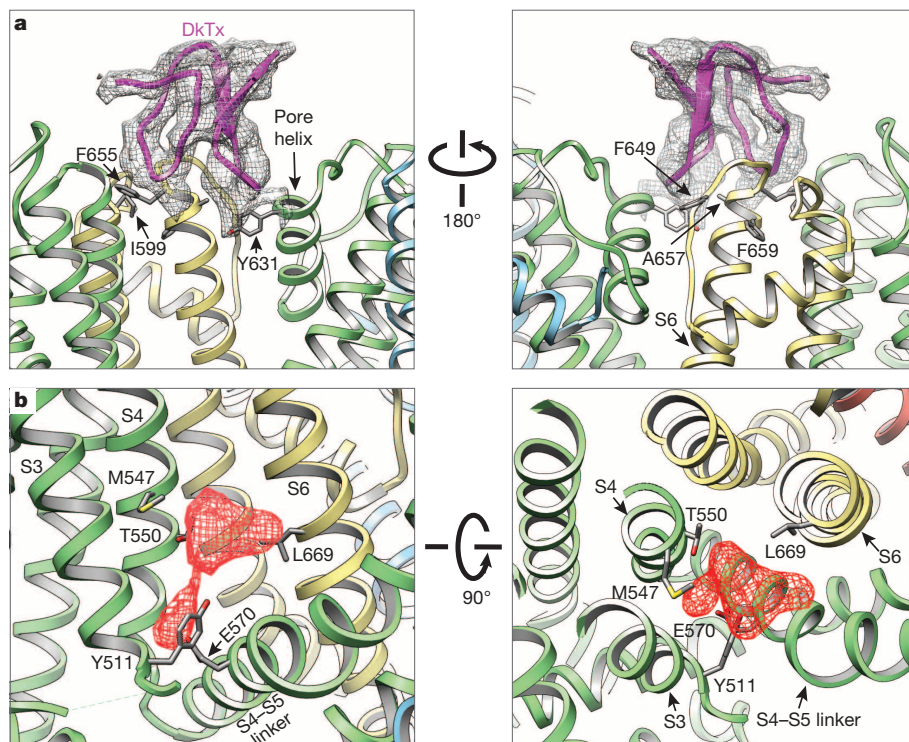


Figure 2 | Binding sites for spider toxin and vanilloid agonists. **a,** Detailed views of interaction between TRPV1 subunits (yellow and green) and one inhibitor cysteine knot from DkTx spider toxin (purple). TRPV1 residues in close proximity to the toxin are highlighted, including four (I599, F649, A657, F659) that, when mutated, render the channel specifically insensitive to DkTx. **b,** Vanilloid-binding pocket defined by EM density of RTX (red, filtered at 4.5 Å with a temperature factor of -200 Å^2 , 8σ) viewed from the side (left) or top-down (that is, from the extracellular face; right). Residues in close proximity to observed densities are highlighted, including several (Y511, M547, T550) that have been previously implicated in vanilloid binding.

the voltage sensor into gating of the pore^{9,11}. Thus, the putative vanilloid site is well positioned to affect channel gating by impinging on two regions (that is, the S4–S5 linker and S6) that probably influence structure of the pore domain (see below). This model is further supported by the capsaicin-bound TRPV1 structure, in which we also see density in the vanilloid pocket (Extended Data Fig. 7). The weaker signal observed with capsaicin may reflect its substantially lower affinity for TRPV1 and smaller mass (molecular weight 305) compared to RTX (molecular weight 628). Interestingly, capsaicin and RTX densities are in close proximity, consistent with overlapping, but non-identical, binding sites²⁶. Intriguingly, we observed density in this same hydrophobic pocket in the apo TRPV1 structure, perhaps representing a lipid or other small hydrophobic molecule associated with the channel in the absence of a competing exogenous ligand (Extended Data Fig. 7). Although our model lacks sufficient detail to reveal precisely how vanilloids bind, we see that Y511 assumes two distinct rotamers in apo versus liganded structures wherein its side chain points away from or into the binding pocket, respectively (Extended Data Fig. 7), suggesting that vanilloid binding involves an ‘induced fit’ mechanism akin to some enzyme–substrate interactions. Taken together, these densities define an important allosteric regulatory site for inflammatory agents, such as anandamide and other endovanilloids, or inverse agonists such as capsazepine.

Pore profiles reveal a dual gate

The three TRPV1 structures described here and in the accompanying study⁸ represent the channel in distinct states that are readily discernible by comparing diameters of their ion conduction pathways (Fig. 3 and Extended Data Fig. 8a–c). In the apo state, the pathway is constricted at the selectivity filter, as well as the lower gate. In the capsaicin-bound structure, there is no change in the selectivity filter, whereas the lower gate is markedly expanded. In the RTX/DkTx structure, the channel is fully open with the ion conduction pathway relieved of any constrictions. These profiles support a dual gating mechanism involving substantial conformational changes at both the selectivity filter and lower gate. TRPV1 has been shown to undergo pore dilation, enabling conduction of organic cations following prolonged activation²⁷. We do not

believe that any of the configurations described here correspond to a dilated state because their pores are probably too narrow to accommodate such large cations. Nonetheless, the plasticity that we observe at both upper and lower gates may allow TRPV1, and possibly other TRPs, to assume a pore-dilated state.

Rearrangements in the outer pore region

Superimposition of TRPV1 structures in apo versus RTX/DkTx-bound states shows a pronounced shift in the relative position of the pore helix. This rearrangement is characterized by a rigid body tilt of the helix away from the central axis of the channel by ~ 1.9 Å and widening of the selectivity filter at its narrowest point, increasing the distance between G643 carbonyl oxygens from 4.6 to 7.6 Å (Fig. 4a–c). The downward movement of M644 also contributes to widening of the filter, where the distance between side chains increases from 5.9 to 13.0 Å (Extended Data Fig. 8a–c). Moreover, conformational changes in the selectivity filter shorten the distance between C α atoms of diagonally apposed D646 residues from 15.6 to 13.0 Å, placing their side chains in a better geometry to coordinate and partially dehydrate cations for efficient conduction. The pore helix shift is also associated with rearrangements in outer pore loops that follow S5 and precede S6 (Fig. 4c), where movements have been suggested to have a role in the activation of TRPV channels by chemical or thermal stimuli^{19,28–31}. For example, two glutamates (E600 and E648) within these loops of TRPV1 are known to be important for acid-evoked potentiation or activation. In the apo TRPV1 structure, these loops are bridged through potential hydrogen bonding between the side chain of E600 and main-chain nitrogen atoms of Y653 and D654 (Fig. 4d). These interactions, which probably maintain the outer pore in a non-conductive conformation in the resting state, are broken in the RTX/DkTx structure owing to rearrangements that increase distance between these loops. We suggest that protonation at E600 similarly disrupts these hydrogen bonds to facilitate gating-associated movements of the pore helix and widening of the selectivity filter as a mechanism by which extracellular protons sensitize the channel to thermal or chemical stimuli.

Generally speaking, the movements we observe in the outer pore and pore helix of TRPV1 stand in contrast to what is seen with potassium

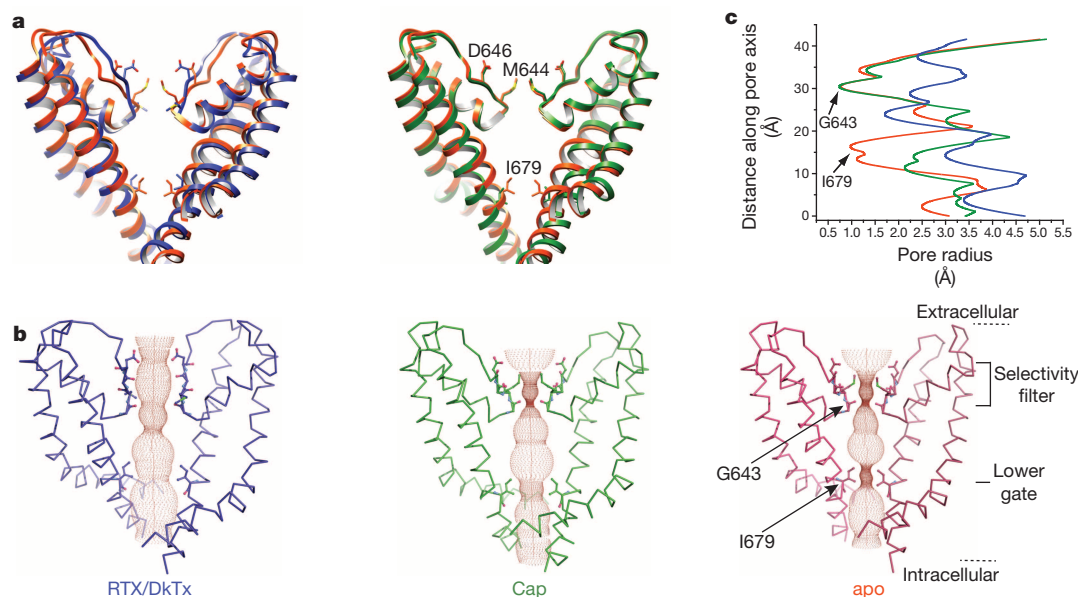


Figure 3 | Comparison of ion permeation pathway in apo versus liganded channels. **a**, Superimposition of S5–P–S6 pore module from apo (orange) versus RTX/DkTx (blue; left)– or capsaicin (green; right)–bound structures. In each case, only two diagonally opposed subunits are shown for clarity. Key residues in the selectivity filter and lower gate are highlighted to display side-chain movements associated with gating. **b**, Solvent-accessible pathway

along the pore mapped using the HOLE program for RTX/DkTx-bound, capsaicin-bound and apo TRPV1 structures. Residues located at the selectivity filter and lower gate are rendered as sticks. **c**, Comparison of pore radii (calculated with the program HOLE) for RTX/DkTx-bound (blue), capsaicin-bound (green) and apo (orange) TRPV1 structures.

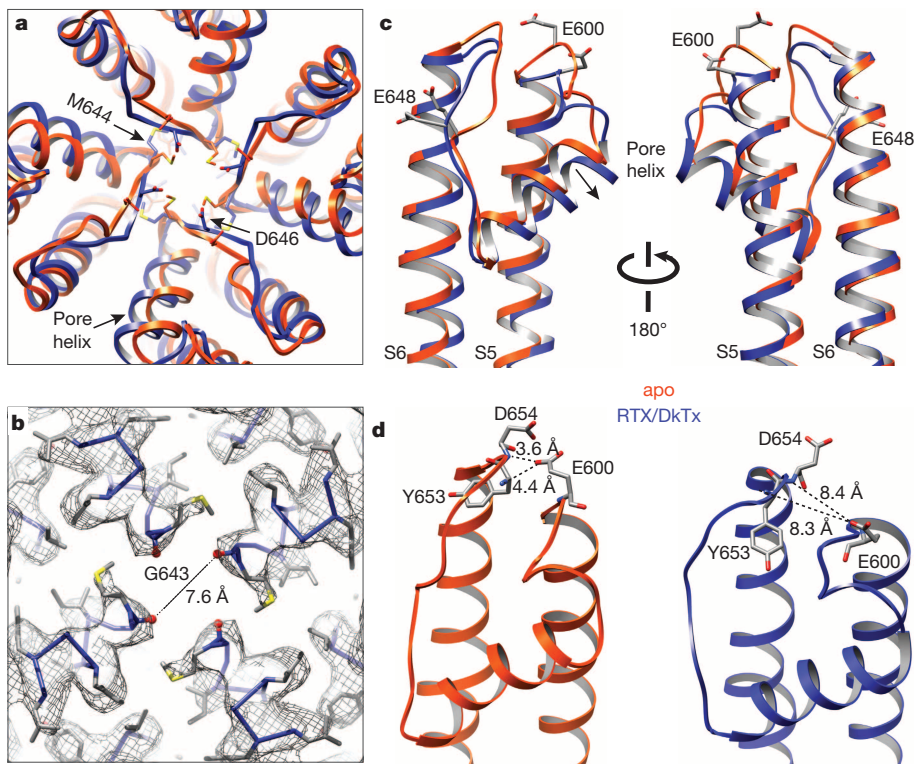


Figure 4 | Structural rearrangements in the outer pore region. **a**, Superimposed top-down views of outer pore regions from apo and RTX/DkTx-bound TRPV1 structures (orange and blue, respectively). Note shift in relative positions of pore helix and selectivity filter. **b**, Density map with atomic model showing distance between diagonally opposed G643 residues, which represents the narrowest point in the outer pore region. **c**, Superimposed side views of the outer pore domains of apo and RTX/DkTx-bound TRPV1 structures (orange and blue, respectively). Residues important for proton-mediated sensitization (E600) or activation (E648) are labelled. Arrow indicates downward rigid body tilt of pore helix in RTX/DkTx structure. **d**, Proximity of E600 to neighbouring residues in the apo structure (left) allows for hydrogen bonding, which is disrupted by rearrangements in the RTX/DkTx-bound structure (right).

channels, where these regions are believed to remain relatively stationary during normal gating^{10,20}. However, there are instances, such as C-type inactivation, ‘flicker transitions’ or multiple subconductance states, that probably result from rearrangements within the K⁺ channel selectivity filter^{32–34}. Indeed, activation of TRPV1 by most stimuli (chemical or thermal) produces flickery behaviour in which channels make frequent excursions to the closed state^{35–37}, perhaps reflecting the dynamic nature of the outer pore. Such flickers are substantially reduced when open channels bind DkTx, presumably reflecting toxin-mediated stabilization of the outer pore (including pore helix and selectivity filter) in fully conducting conformations¹⁵.

The structure of the TRPV1 outer pore in the presence of capsaicin is superimposable with that of the apo, closed channel (Extended Data Fig. 8d, e). In the absence of DkTx, this structure is unlikely to be trapped in its fully open state and may, instead, represent the channel during one of its frequent excursions to partially or transiently closed states. Nonetheless, the channel has undergone conformational changes associated with gating, including movements of the S4–S5 linker and lower gate, as discussed below. Thus, we conclude that our capsaicin–TRPV1 structure represents a partially activated state.

Opening of the lower gate

As we show in the accompanying study⁸, I679 in S6 forms a hydrophobic seal sufficiently narrow (5.3 Å) to block permeation by hydrated ions when TRPV1 is in its closed conformation. In the RTX/DkTx-bound structure, we see a remarkable disruption of this seal with consequent expansion of the lower gate to 9.3 Å in diameter (Fig. 5a, b). This is driven by conformational changes in the lower half of the S6 helix, resulting in ~30° rotation of I679 side chains away from the central canal. With capsaicin, the lower gate expands to 7.6 Å (Fig. 5c, d). These changes in pore diameter are associated with a slight bowing out of S6 away from the vertical axis, which occurs in nearly opposite directions for capsaicin versus RTX/DkTx (Extended Data Fig. 9). Consequently, rotation of I679 away from the central axis is more pronounced for RTX/DkTx, accounting for wider opening of the lower gate compared to capsaicin (Extended Data Fig. 8).

In VGICs, movement within the S1–S4 voltage sensor is coupled to channel opening through the S4–S5 linker helix^{9,10}. In K_v channels,

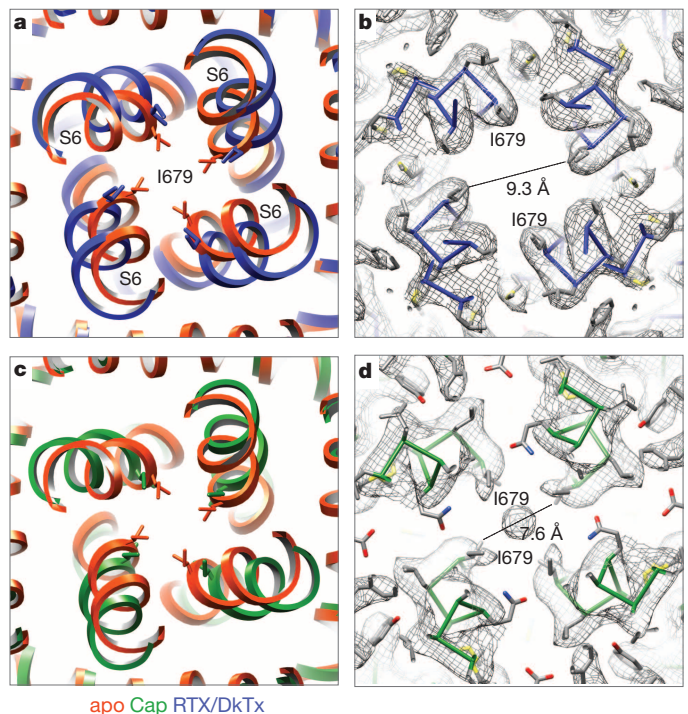


Figure 5 | Opening of the lower gate. **a**, Superimposition of inner pore region from apo and RTX/DkTx-bound TRPV1 structures (orange and blue, respectively). The residue (I679) that forms the hydrophobic seal of the lower gate is highlighted, and its side chain shown in stick format. **b**, Density map with atomic model showing distances between I679 in the RTX/DkTx-bound channel. Note substantial expansion of the lower gate relative to apo structure. **c**, Superimposition of apo and capsaicin-bound TRPV1 inner pore regions (orange and green, respectively). **d**, Density map with atomic model showing distances between I679 in the capsaicin-bound channel. Note expansion of the lower gate relative to apo-structure, but to a lesser extent than seen in RTX/DkTx-bound channel. The density at the central cavity may represent noise amplified by applying symmetry, or a trapped hydrated ion. Interestingly, this density is not observed in apo or RTX/DkTx structures.

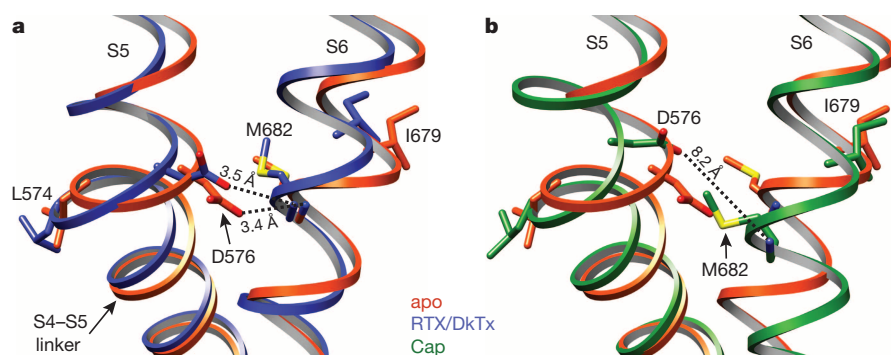


Figure 6 | Coupling of S4-S5 linker and S6 helix. **a, b,** Superimposition of apo (orange) with RTX/DkTx (blue; **a**) or capsaicin (green; **b**)-bound structures highlighting interactions between residues in the S4-S5 linker and S6 helix. Through these interactions, movement of the S4-S5 linker is translated to rotation or displacement of I679 away from the central axis, breaking the hydrophobic seal to open the lower gate.

lateral movement of the S4-S5 linker leads to splaying of the lower S6 segment below a kink formed by a conserved -Pro-X-Pro- motif or glycine residue, resulting in widening of the activation gate at the apex of the inverted teepee^{10,38}. In TRPV1, the S4-S5 linker and S6 also interact, but in a way that suggests a distinct gating mechanism. Whereas in VGICs these contacts take place well above the activation gate, in TRPV1 they occur just slightly beneath I679, placing them in direct alignment with the lower gate (Fig. 6). Thus, in both apo and RTX/DkTx structures, hydrogen bonding is observed between two residues that are invariant in all TRPV subtypes, namely D576 in the S4-S5 linker and M682 in S6. Consequently, movement at the junction between S5 and the S4-S5 linker (~ 2 Å) produces a concomitant shift in S6 and associated movement of I679 away from the central canal. In the capsaicin-bound structure, we see a somewhat larger shift (~ 2.7 Å) of the S4-S5 linker (with loss of hydrogen bonding between D576 and M682), and outward movement of I679 (Fig. 6b). Thus, we propose that opening of the TRPV1 lower gate involves more subtle rearrangements in S6 compared to VGICs, and occurs closer to the point of interaction with the S4-S5 linker. Indeed, the S6 helix in TRPV1 lacks a proline or glycine kink that underlies the more pronounced outward splaying mechanism of K_V gate opening¹⁰. Movement of the S4-S5 linker and S6 is also accompanied by lateral displacement of the TRP domain at the helical 'elbow' proximal to S6 (Extended Data Fig. 9c, d), where a 3.5 Å displacement between C α atoms of I696 is readily seen when comparing apo versus RTX/DkTx-bound structures. These observations are consistent with the idea that the TRP domain functions as an allosteric modulatory site³⁹.

Static nature of the S1-S4 domain

In contrast to VGICs, we find that the S1-S4 domain of TRPV1 is virtually superimposable when comparing apo and activated structures (Extended Data Fig. 9), suggesting that this domain has a less active role in TRP channel gating. This may explain why many gain-of-function mutations in TRPV subtypes are found within the S4-S5 linker, S5-P-S6 pore and TRP domain, but not within the S1-S4 domain^{18,40,41}. Taken together, we posit that S1-S4 domain functions as a passive anchor upon which the S4-S5 linker moves in response to ligand binding. This is consistent with our observation that opening of the TRPV1 inner gate does not involve splaying of the lower half of S6, which would be constrained by the static S1-S4 domain.

TRPV1 and certain other TRP subtypes exhibit weak voltage dependence⁴²⁻⁴⁴, for which the structural basis remains obscure. Our observations suggest that this is unlikely to involve movement within S1-S4, consistent with the fact that TRPV1 lacks positive charges in S4 that are characteristic of VGICs. However, as detailed above, the outer pore undergoes substantial downward movement during gating, away from the extracellular face of the bilayer. Because this helix forms a negative dipole at its carboxy-terminal end, movement should be facilitated during membrane depolarization, enhancing open probability at positive membrane potentials, especially when the lower gate is open and membrane potential is focused on the upper gate. This conformational change may contribute to aspects of voltage sensitivity, such as

outward rectification characteristic of many TRP channel subtypes⁴⁴. This idea (and that of dual gating) is consistent with TRPV4 mutagenesis studies suggesting that rectification is mediated by a voltage-dependent gating mechanism that operates in tandem with the primary intracellular gate⁴⁵.

Summary

TRP channels are noted for their ability to function as polymodal signal integrators³⁷, which in the case of TRPV1 underlies its role in pain hypersensitivity¹. The structures presented here demonstrate that regions targeted by inflammatory agents induce or undergo substantial conformational changes associated with gating. These include the outer pore domain, which we show to be unusually dynamic compared to VGICs, as well as the hydrophobic pocket defined by the external surface of the S3-S4 helices, S4-S5 linker and S6 helix. Thus, proalgesic agents can modulate channel activity by exerting immediate effects on either of two restriction points defined by the selectivity filter and lower gate. Although these regions can be targeted independently, pharmacological observations suggest that they are allosterically coupled. For example, distinct stimuli (chemical or thermal) show cross sensitization, and the inverse vanilloid agonist capsazepine inhibits channel activation by capsaicin, protons, spider toxins, heat or constitutive pore helix mutations^{18,37,46,47}. We suspect that the pore helix represents a critical structural element in coupling upper and lower gates; in the RTX/DkTx-bound structure, downward tilt of the pore helix away from the central canal occurs concomitantly with movement of S5, probably through close side-chain interactions that foster physical coupling between the two helices to facilitate opening of the lower gate (Extended Data Fig. 10a, b). Such dynamic communication between the upper and lower gates could underlie integration of diverse physiological signals (Extended Data Fig. 10c).

A fascinating aspect of TRPV1 function relates to its robust thermosensitivity. Heat is ubiquitous and can affect the energy landscape of both the upper and lower gates. Indeed, structure-function and pharmacological studies have implicated various regions as determinants of thermosensitivity, including cytoplasmic and outer pore domains^{17,19,48-50}. Further insights into the mechanism of heat-evoked gating would be greatly facilitated by obtaining 'snapshots' of the channel at different temperatures. We believe that recent advances in cryo-EM make it possible to achieve this goal, which is probably out of reach for more established methods of protein crystallography.

METHODS SUMMARY

A minimal functional rat TRPV1 construct was engineered, expressed and functionally analysed as described in the accompanying study⁸. For preparation of channel in complex with different pharmacological agents, TRPV1 protein in amphipols was mixed with DkTx and RTX or capsaicin for 30 min before grid preparation. Single-particle cryo-EM, including data acquisition and image processing, were carried out as described in the accompanying paper. For model building, the apo structure of TRPV1 was docked into the 3D density maps of capsaicin- and RTX/DkTx-bound structures as a rigid body. The S1-S4 domain fits nicely into both maps, therefore only slight adjustments were performed at this part of the model. On the other hand, the S4-S5 linker, pore module (S5-P-S6) and TRP

domain of the TRPV1 apo structure exhibit substantial deviation from both maps, indicating structural rearrangements during ligand binding. This part of the model was rebuilt in Coot using the apo structure as a reference. The density map for the pore module of the capsaicin-bound structure is of sufficient quality to assign most side chains in this region, allowing us to obtain an accurate pore profile by calculating pore radii using HOLE program. The map of RTX/DkTx-bound structure is of sufficient quality to build the model without ambiguity. A polyalanine model based on the NMR structure of hanatoxin was docked onto the DkTx density with slight adjustment.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 July; accepted 30 October 2013.

- Basbaum, A. I., Bautista, D. M., Scherrer, G. & Julius, D. Cellular and molecular mechanisms of pain. *Cell* **139**, 267–284 (2009).
- Baconguis, I. & Gouaux, E. Structural plasticity and dynamic selectivity of acid-sensing ion channel-spider toxin complexes. *Nature* **489**, 400–405 (2012).
- Hansen, S. B., Tao, X. & MacKinnon, R. Structural basis of PIP2 activation of the classical inward rectifier K⁺ channel Kir2.2. *Nature* **477**, 495–498 (2011).
- Hattori, M. & Gouaux, E. Molecular mechanism of ATP binding and ion channel activation in P2X receptors. *Nature* **485**, 207–212 (2012).
- Whorton, M. R. & MacKinnon, R. X-ray structure of the mammalian GIRK2- β - γ -G-protein complex. *Nature* **498**, 190–197 (2013).
- Bohlen, C. J. & Julius, D. Receptor-targeting mechanisms of pain-causing toxins: how ow? *Toxicon* **60**, 254–264 (2012).
- Vriens, J., Appendino, G. & Nilius, B. Pharmacology of vanilloid transient receptor potential cation channels. *Mol. Pharmacol.* **75**, 1262–1279 (2009).
- Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* <http://dx.doi.org/10.1038/nature12823> (this issue).
- Catterall, W. A. Ion channel voltage sensors: structure, function, and pathophysiology. *Neuron* **67**, 915–928 (2010).
- Long, S. B., Campbell, E. B. & MacKinnon, R. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* **309**, 903–908 (2005).
- Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **456**, 891–897 (2008).
- Zhu, S., Darbon, H., Dyason, K., Verdonck, F. & Tytgat, J. Evolutionary origin of inhibitor cystine knot peptides. *FASEB J.* **17**, 1765–1767 (2003).
- Phillips, L. R. et al. Voltage-sensor activation with a tarantula toxin as cargo. *Nature* **436**, 857–860 (2005).
- Swartz, K. J. & MacKinnon, R. Hanatoxin modifies the gating of a voltage-dependent K⁺ channel through multiple binding sites. *Neuron* **18**, 665–673 (1997).
- Bohlen, C. J. et al. A bivalent tarantula toxin activates the capsaicin receptor, TRPV1, by targeting the outer pore domain. *Cell* **141**, 834–845 (2010).
- Siemens, J. et al. Spider toxins activate the capsaicin receptor to produce inflammatory pain. *Nature* **444**, 208–212 (2006).
- Grandl, J. et al. Temperature-induced opening of TRPV1 ion channel is stabilized by the pore domain. *Nature Neurosci.* **13**, 708–714 (2010).
- Myers, B. R., Bohlen, C. J. & Julius, D. A yeast genetic screen reveals a critical role for the pore helix domain in TRP channel gating. *Neuron* **58**, 362–373 (2008).
- Yang, F., Cui, Y., Wang, K. & Zheng, J. Thermosensitive TRP channel pore turret is part of the temperature activation pathway. *Proc. Natl Acad. Sci. USA* **107**, 7083–7088 (2010).
- Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K⁺ channel–Fab complex at 2.0 Å resolution. *Nature* **414**, 43–48 (2001).
- Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
- Chou, M. Z., Mtui, T., Gao, Y. D., Kohler, M. & Middleton, R. E. Resiniferatoxin binds to the capsaicin receptor (TRPV1) near the extracellular side of the S4 transmembrane domain. *Biochemistry* **43**, 2501–2511 (2004).
- Gavva, N. R. et al. Molecular determinants of vanilloid sensitivity in TRPV1. *J. Biol. Chem.* **279**, 20283–20295 (2004).
- Jordt, S. E. & Julius, D. Molecular basis for species-specific sensitivity to “hot” chili peppers. *Cell* **108**, 421–430 (2002).
- Phillips, E., Reeve, A., Bevan, S. & McIntyre, P. Identification of species-specific determinants of the action of the antagonist capsazepine and the agonist PPAHV on TRPV1. *J. Biol. Chem.* **279**, 17165–17172 (2004).
- Szallasi, A., Blumberg, P. M., Annicelli, L. L., Krause, J. E. & Cortright, D. N. The cloned rat vanilloid receptor VR1 mediates both R-type binding and C-type calcium response in dorsal root ganglion neurons. *Mol. Pharmacol.* **56**, 581–587 (1999).
- Chung, M. K., Guler, A. D. & Caterina, M. J. TRPV1 shows dynamic ionic selectivity during agonist stimulation. *Nature Neurosci.* **11**, 555–564 (2008).
- Jordt, S. E., Tominaga, M. & Julius, D. Acid potentiation of the capsaicin receptor determined by a key extracellular site. *Proc. Natl Acad. Sci. USA* **97**, 8134–8139 (2000).
- Kim, S. E., Patapoutian, A. & Grandl, J. Single residues in the outer pore of TRPV1 and TRPV3 have temperature-dependent conformations. *PLoS ONE* **8**, e59593 (2013).
- Ryu, S., Liu, B., Yao, J., Fu, Q. & Qin, F. Uncoupling proton activation of vanilloid receptor TRPV1. *J. Neurosci.* **27**, 12797–12807 (2007).
- Yeh, B. I., Kim, Y. K., Jabbar, W. & Huang, C. L. Conformational changes of pore helix coupled to gating of TRPV5 by protons. *EMBO J.* **24**, 3224–3234 (2005).
- Bernèche, S. & Roux, B. A gate in the selectivity filter of potassium channels. *Structure* **13**, 591–600 (2005).
- Cuello, L. G., Jogini, V., Cortes, D. M. & Perozo, E. Structural mechanism of C-type inactivation in K⁺ channels. *Nature* **466**, 203–208 (2010).
- Hoshi, T. & Armstrong, C. M. C-type inactivation of voltage-gated K⁺ channels: pore constriction or dilation? *J. Gen. Physiol.* **141**, 151–160 (2013).
- Hui, K., Liu, B. & Qin, F. Capsaicin activation of the pain receptor, VR1: multiple open states from both partial and full binding. *Biophys. J.* **84**, 2957–2968 (2003).
- Liu, B., Hui, K. & Qin, F. Thermodynamics of heat activation of single capsaicin ion channels VR1. *Biophys. J.* **85**, 2988–3006 (2003).
- Tominaga, M. et al. The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
- Jiang, Y. et al. The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
- Latorre, R., Zaelzer, C. & Brauchi, S. Structure–functional intimacies of transient receptor potential channels. *Q. Rev. Biophys.* **42**, 201–246 (2009).
- Dai, J. et al. TRPV4-pathway, a novel channelopathy affecting diverse systems. *J. Hum. Genet.* **55**, 400–402 (2010).
- Lin, Z. et al. Exome sequencing reveals mutations in TRPV3 as a cause of Olmsted syndrome. *Am. J. Hum. Genet.* **90**, 558–564 (2012).
- Latorre, R., Vargas, G., Orta, G. & Brauchi, S. in *TRP Ion Channel Function in Sensory Transduction and Cellular Signaling Cascades* (Frontiers in Neuroscience) (eds Liedtke, W. B. & Heller, S.) (2007).
- Matta, J. A. & Ahern, G. P. Voltage is a partial activator of rat thermosensitive TRP channels. *J. Physiol. (Lond.)* **585**, 469–482 (2007).
- Nilius, B. et al. Gating of TRP channels: a voltage connection? *J. Physiol. (Lond.)* **567**, 35–44 (2005).
- Loukin, S., Su, Z., Zhou, X. & Kung, C. Forward genetic analysis reveals multiple gating mechanisms of TRPV4. *J. Biol. Chem.* **285**, 19884–19890 (2010).
- Cao, E., Cordero-Morales, J. F., Liu, B., Qin, F. & Julius, D. TRPV1 channels are intrinsically heat sensitive and negatively regulated by phosphoinositide lipids. *Neuron* **77**, 667–679 (2013).
- van der Stelt, M. & Di Marzo, V. Endovanilloids. Putative endogenous ligands of transient receptor potential vanilloid 1 channels. *Eur. J. Biochem.* **271**, 1827–1834 (2004).
- Brauchi, S., Orta, G., Salazar, M., Rosenmann, E. & Latorre, R. A hot-sensing cold receptor: C-terminal domain determines thermosensation in transient receptor potential channels. *J. Neurosci.* **26**, 4835–4840 (2006).
- Papakosta, M. et al. The chimeric approach reveals that differences in the TRPV1 pore domain determine species-specific sensitivity to block of heat activation. *J. Biol. Chem.* **286**, 39663–39672 (2011).
- Yao, J., Liu, B. & Qin, F. Modular thermal sensors in temperature-gated transient receptor potential (TRP) channels. *Proc. Natl Acad. Sci. USA* **108**, 11109–11114 (2011).

Acknowledgements We thank X. Li for assistant in data acquisition using TF30 Polara and K2 Summit camera, and J.P. Armache, C. Bohlen, J. Cordero-Morales and J. Osteen for discussion and reading of the manuscript. This work was supported by grants from the National Institutes of Health (R01GM098672 and S10RR026814 to Y.C. and R01NS065071 and R01NS047723 to D.J.), the National Science Foundation (DBI-0960271 to D. Agard and Y.C.) and the University of California, San Francisco Program for Breakthrough Biomedical Research (Y.C.). E.C. was a fellow of the Damon Runyon Cancer Research Foundation.

Author Contributions All authors designed experiments. E.C. expressed and purified all protein samples used in this work and performed all functional studies. M.L. carried out all cryo-EM experiments, including data acquisition and processing. E.C. built the atomic model on the basis of cryo-EM maps. All authors analysed data and wrote the manuscript.

Author Information 3D cryo-EM density maps of TRPV1 complexes without low-pass filter and amplitude modification have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-5776 (TRPV1–RTX/DkTx) and EMD-5777 (TRPV1–capsaicin). The coordinates of atomic models of TRPV1 in these two states have been deposited in the Protein Data Bank under the accession numbers 3J5Q and 3J5R. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J. (david.julius@ucsf.edu) or Y.C. (yicheng@ucsf.edu).

METHODS

Sample preparation and data acquisition for cryo-EM analysis. TRPV1 protein was purified as described in the accompanying study⁸. For preparation of channel in complex with different pharmacological agents, TRPV1 protein (final concentration 5 μ M) in amphipols was mixed with DkTx (final concentration 10 μ M), which has two toxin moieties and a total molecular weight of 8.5 kDa, and RTX (50 μ M; molecular weight 628 Da), or with capsaicin (50 μ M; molecular weight 305 Da) for 30 min before grid preparation. Cryo-EM grids were prepared by the same procedure as described in the accompanying paper⁸. Frozen hydrated TRPV1 complexes were imaged on TF30 Polara microscope operated at 300 kV using K2 Summit in dose fractionation super-resolution counting mode. Imaging conditions were identical with those used for imaging channel alone, as described in the accompanying study⁸.

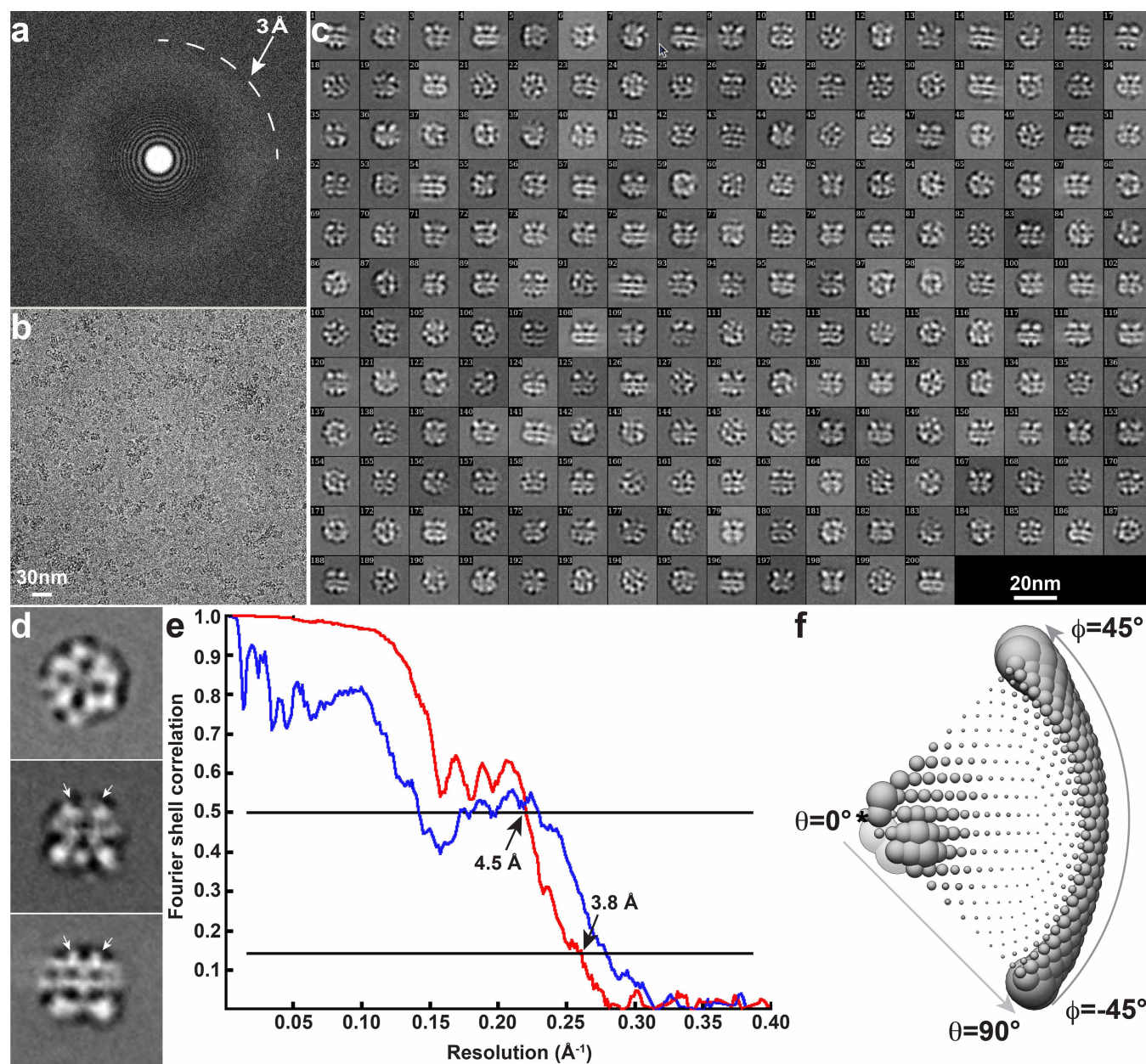
Image processing. Particle picking, contrast transfer function determination and two-dimensional classifications were carried out by the same procedure as described in the accompanying paper⁸. After particle picking and two-dimensional screening, 148,670 particles of RTX/DkTx-bound TRPV1 and 95,897 particles of capsaicin-bound TRPV1 were selected for 3D classification in RELION⁵¹, with the 3D reconstruction of apo TRPV1 filtered to a resolution of 60 Å as an initial model. Refinement of the final selected 3D classes was also carried out in RELION. 36,158 and 33,238 particles of RTX/DkTx- and capsaicin-bound TRPV1, respectively, were used to calculate the final maps. Resolution of final 3D reconstruction was estimated by gold-standard FSC = 0.143 criteria, after applying a soft spherical mask on the two reconstructions independently refined from the half-data sets²¹. The resolutions of the final 3D reconstructions were 3.8 Å for RTX/DkTx-bound TRPV1 and 4.2 Å for capsaicin-bound TRPV1. For model building and visualization, amplitude of the final 3D density map was amplified by a temperature factor of -150 Å². All maps deposited to the Electron Microscopy Data Bank (EMDB) database are the raw maps without amplitude sharpening, masking or filtering.

Before calculating difference maps between 3D reconstructions of liganded and apo TRPV1, all maps were filtered to the same resolution of 6 Å and applying the same temperature factor of -100 Å². Difference maps were calculated by subtracting the density of apo TRPV1 from the liganded TRPV1, using the program 'diffmap.exe' (written by N. Grigorieff). For visualizing ligand densities (Fig. 2b and Extended Data Fig. 7b), liganded and apo TRPV1 maps were filtered to 4.5 Å

with the same temperature factor (-200 Å²) and normalized using MAPMAN (Uppsala Software Factory) and visualized with the same threshold (8σ). UCSF Chimera⁵² was used to visualize and segment the cryo-EM maps, and the rigid body fitting of atomic model was done using the fit-in-map function from Chimera.

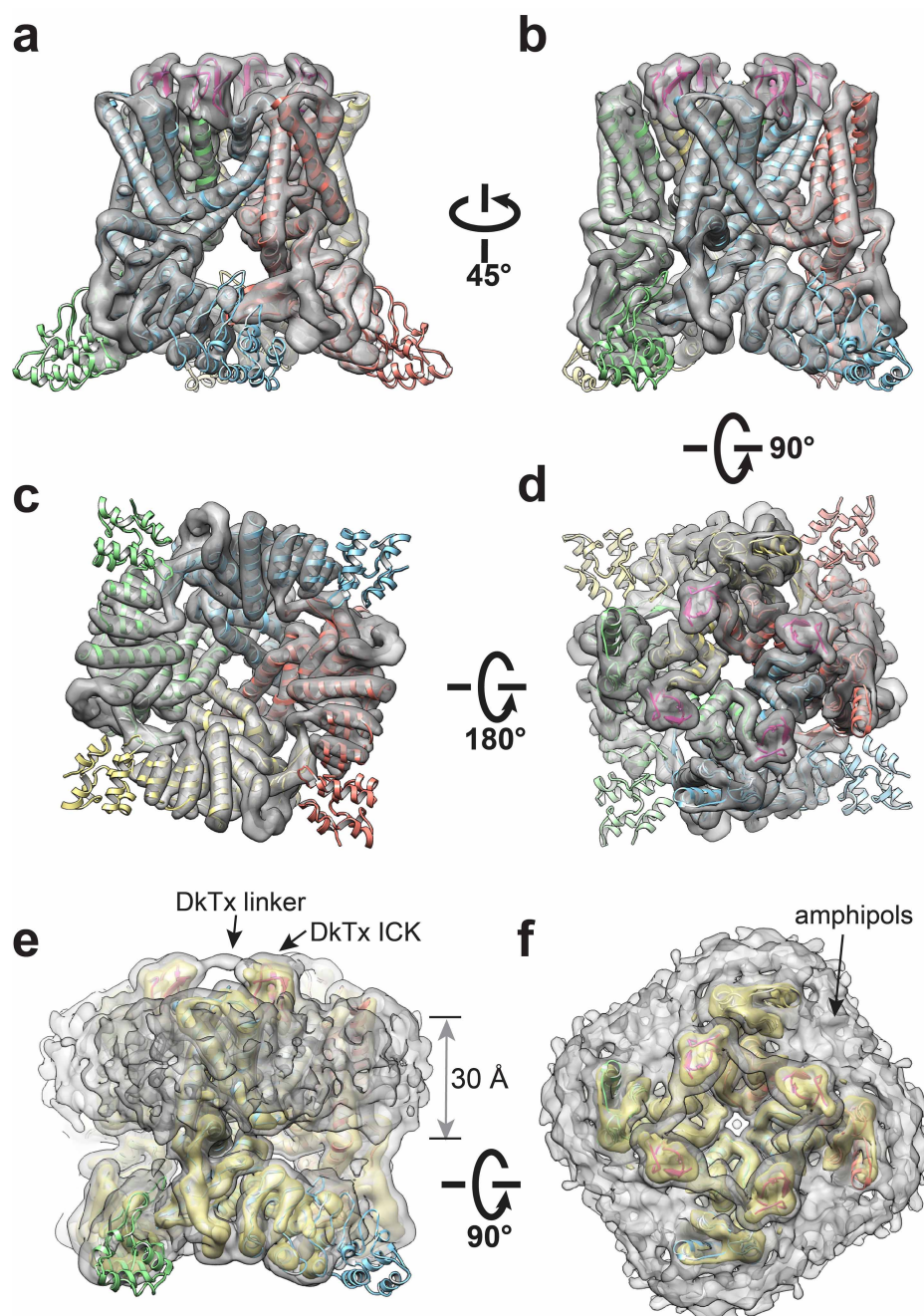
Model building in Coot. The apo structure of TRPV1 was docked into the 3D density maps of capsaicin- and RTX/DkTx-bound structures as a rigid body. The S1–S4 domain fits nicely into both maps, therefore only slight adjustments were performed at this part of the model. On the other hand, the S4–S5 linker, pore module (S5–P–S6) and TRP domain of the TRPV1 apo structure exhibit substantial deviation from both maps, indicating structural rearrangements during ligand binding. This part of the model was rebuilt in Coot^{53,54} using the apo structure as a reference. The density map for the pore module of the capsaicin-bound structure is of sufficient quality to assign most side chains in this region, allowing us to obtain an accurate pore profile by calculating pore radii using HOLE program⁵⁵. However, density for the TRP domain, especially the part that immediately follows S6, does not allow for side-chain resolution, probably reflecting the dynamic nature of this region during gating. Nonetheless, the invariant W697 in helical TRP domain shows excellent density and serves as a landmark to orient the entire helix. The map of the RTX/DkTx-bound structure is of sufficient quality to build the model without ambiguity. A polyaniline model (a total of 31 alanine) based on the NMR structure of hanatoxin (PDB 1D1H)⁵⁶ was docked onto the DkTx density with slight adjustment.

51. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
52. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
53. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
54. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
55. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
56. Takahashi, H. et al. Solution structure of hanatoxin1, a gating modifier of voltage-dependent K⁺ channels: common surface features of gating modifier toxins. *J. Mol. Biol.* **297**, 771–780 (2000).



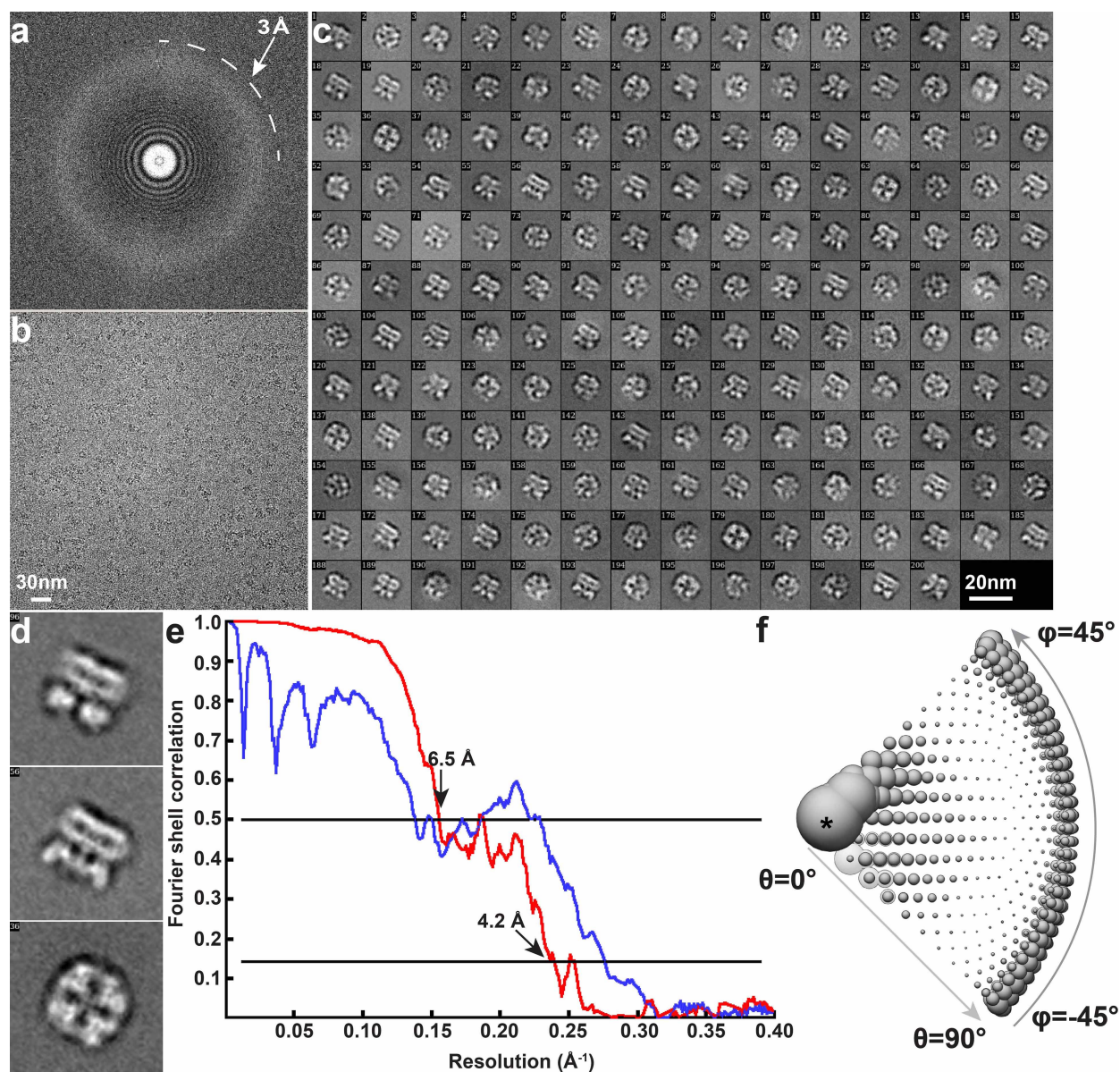
Extended Data Figure 1 | Cryo-EM of TRPV1 in complex with RTX and DkTx. **a**, **b**, Fourier transform (**a**) of a representative image (**b**). **c**, Two dimensional (2D) class averages of cryo-EM particles. **d**, Enlarged view of three representative 2D class averages. Arrows indicate DkTx densities near the channel pore. **e**, Gold-standard FSC curve (red) of the final 3D reconstruction,

marked with the resolutions that correspond to FSC = 0.5 and 0.143. The FSC curve between the final map and that calculated from the atomic model is shown in blue. **f**, Euler angle distribution of all particles used for calculating the final 3D reconstruction. The sizes of balls represent the number of particles. The accuracy of rotation is 5.213°, as reported by RELION.



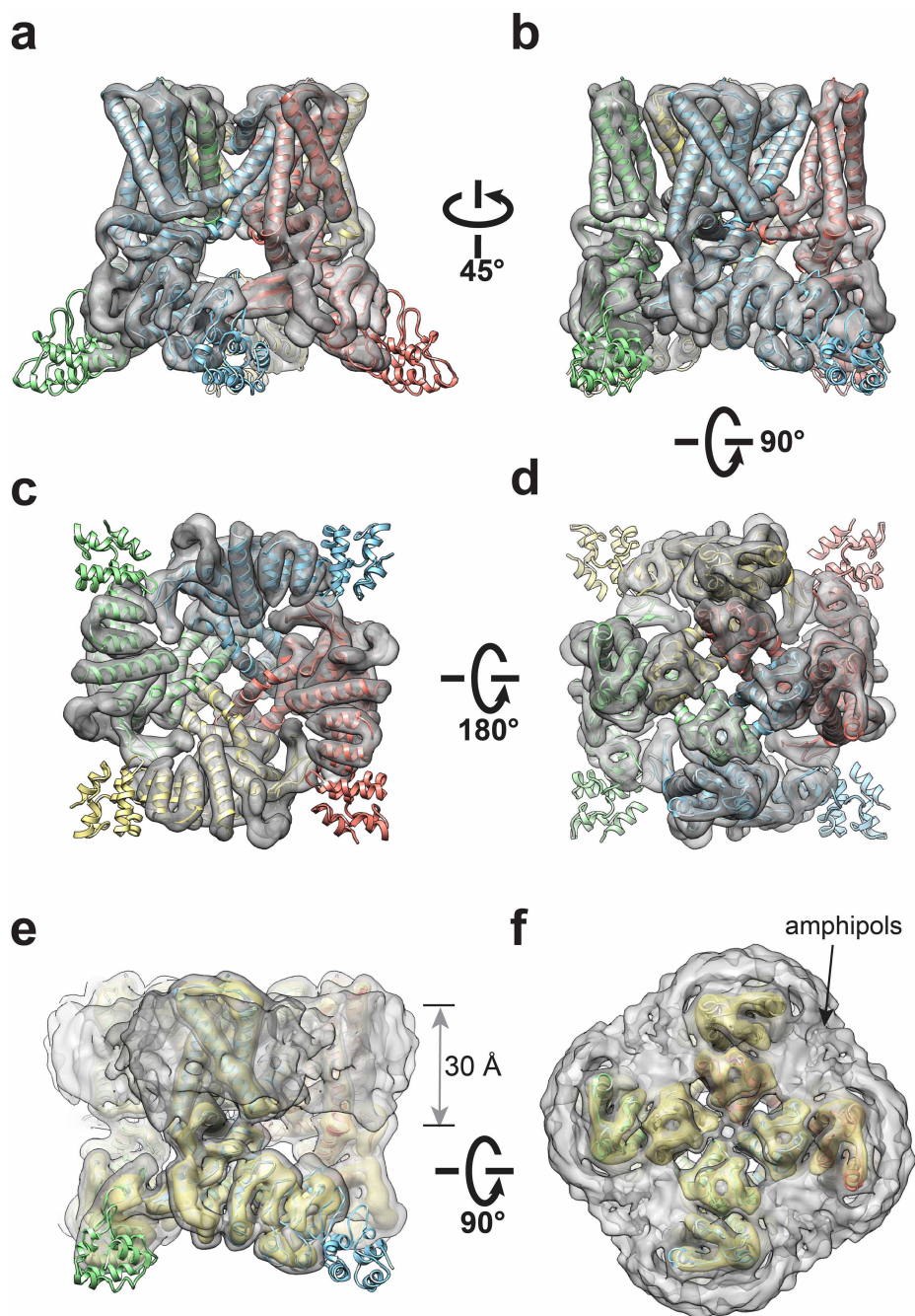
Extended Data Figure 2 | 3D reconstruction of TRPV1-RTX/DkTx complexes filtered at 6 Å resolution. **a–d**, Four different views of the 3D reconstruction low-pass filtered at 6 Å and amplified by a temperature factor of -100 Å^2 , fitted with *de novo* atomic model of TRPV1-RTX/DkTx complex (toxin is shown in magenta) built as described in Methods. **e, f**, Two views of the

3D reconstruction displayed at two different isosurface levels (high in yellow and low in grey). At the low isosurface level, the belt-shaped density of amphipols is visible with a thickness of $\sim 30 \text{ Å}$. DkTx-related densities are also clearly visible, including the linker peptide that connects the toxin's two inhibitor cysteine knot (ICK) moieties, as noted.



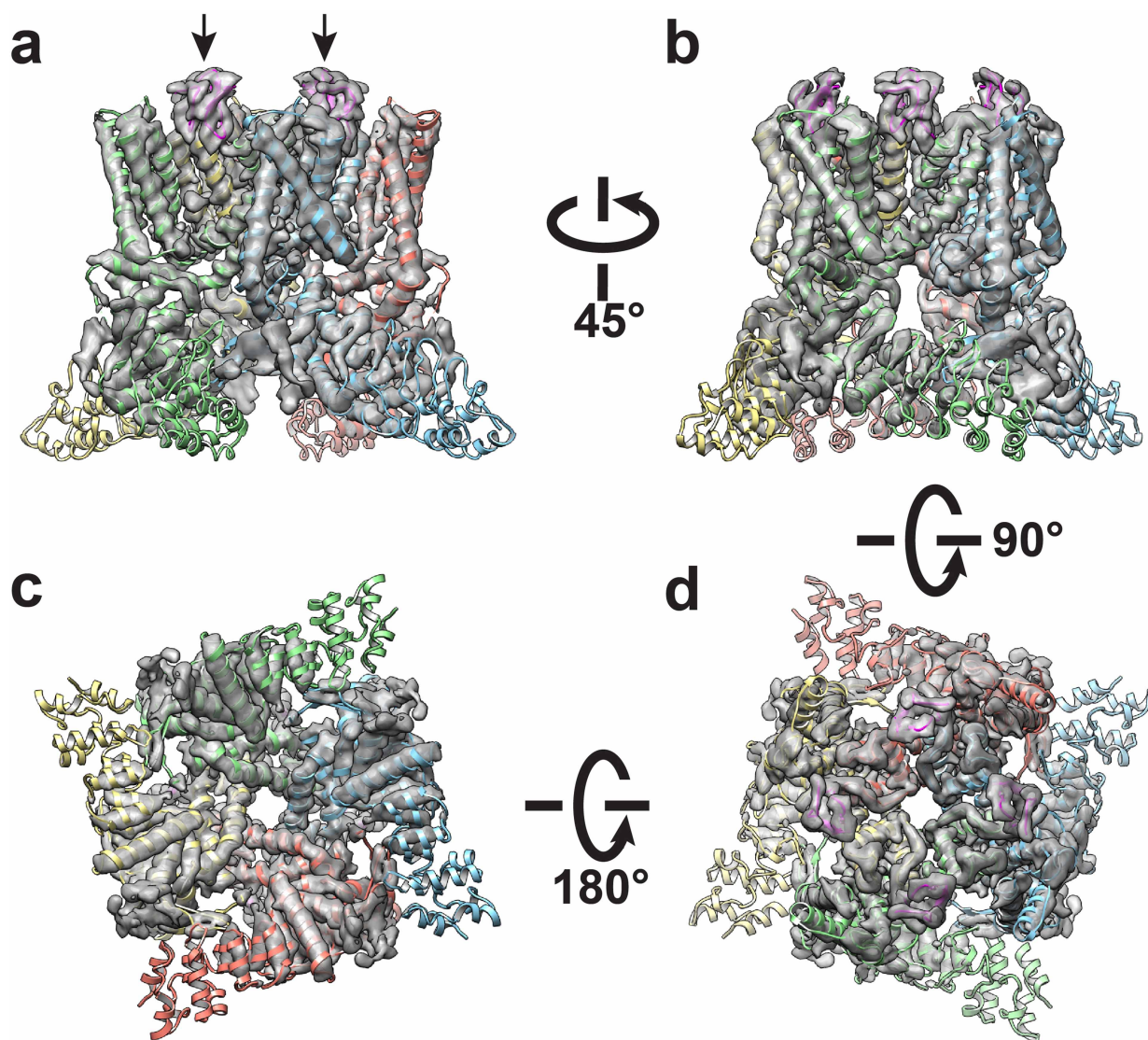
Extended Data Figure 3 | Cryo-EM of TRPV1 in complex with capsaicin.
a, b, Fourier transform (**a**) of a representative image (**b**). **c**, 2D class averages of cryo-EM particles. **d**, Enlarged view of three representative 2D class averages. **e**, Gold-standard FSC curve (red) of the final 3D reconstruction, marked with the resolutions that correspond to FSC = 0.5 and 0.143. The FSC curve

between the final map and that calculated from the atomic model is shown in blue. **f**, Euler angle distribution of all particles used for calculating the final 3D reconstruction. The sizes of balls represent the number of particles. The accuracy of rotation is 4.989°, as reported by RELION.



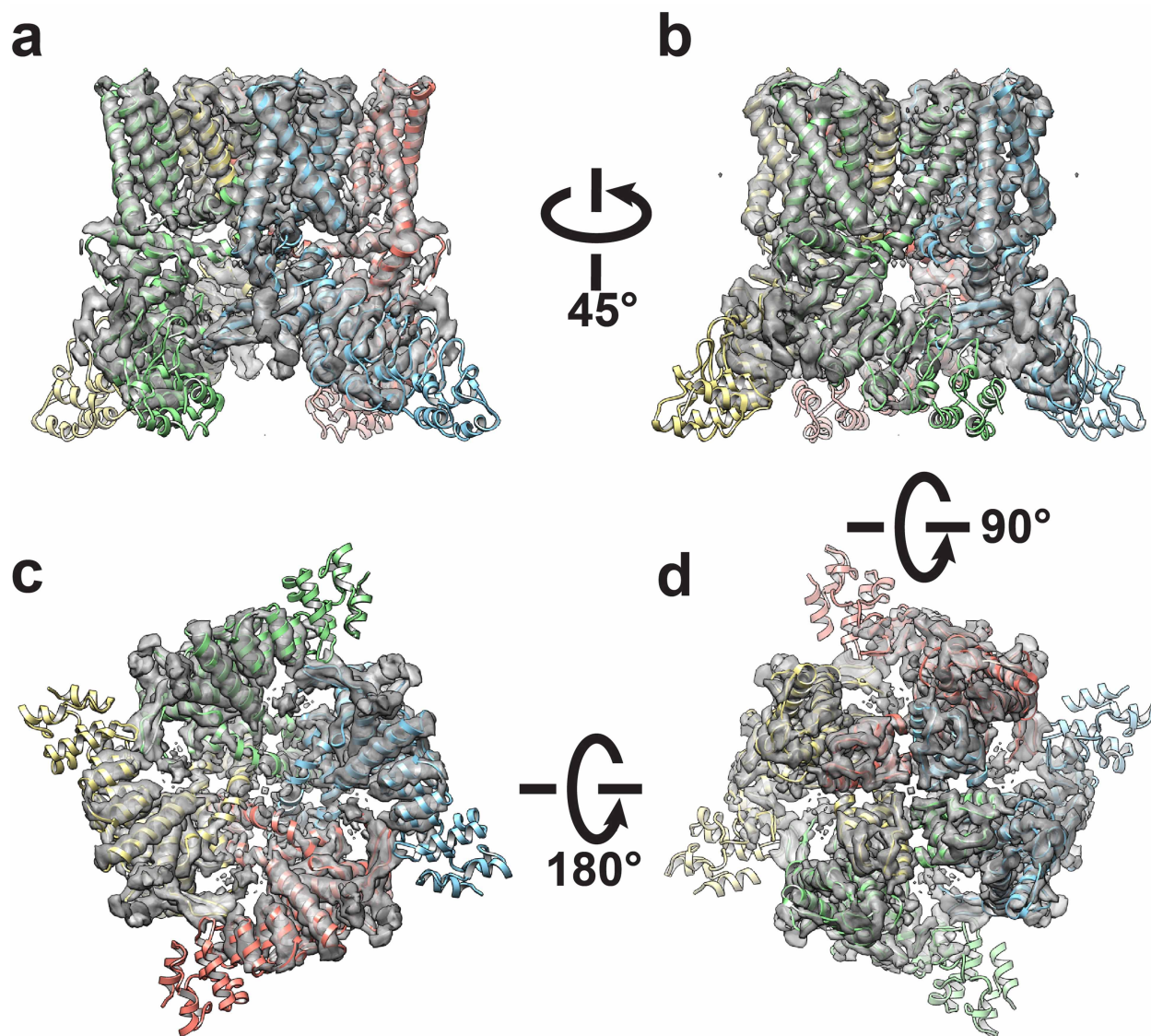
Extended Data Figure 4 | 3D reconstruction of TRPV1-capsaicin complex low-pass filtered at 6 Å resolution. a–d, Four different views of the 3D reconstruction low-pass filtered at 6 Å and amplified by a temperature factor of -100 Å^2 , fitted with *de novo* atomic model of TRPV1-capsaicin complex built

as described in Methods. e, f, Two views of the 3D reconstruction displayed at two different isosurface levels (high in yellow and low in grey). At the low isosurface level, the belt-shaped density of amphipols is visible with a thickness of $\sim 30 \text{ Å}$.



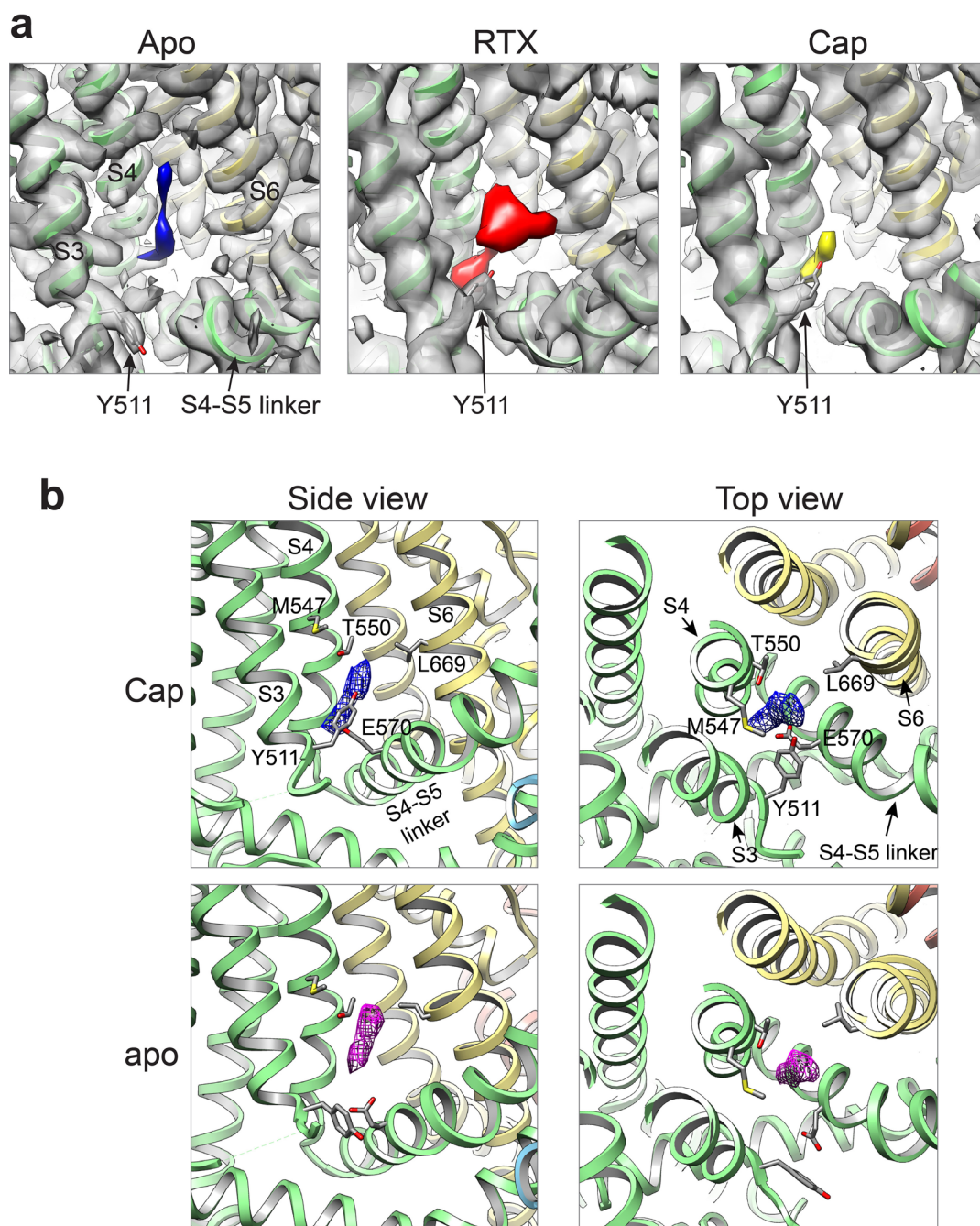
Extended Data Figure 5 | 3D reconstruction of TRPV1-RTX/DkTx complex low-pass filtered at 3.8 Å resolution. a–d, Four different views of the 3D reconstruction low-pass filtered at 3.8 Å with a temperature factor of

-100 Å^2 , fitted with *de novo* atomic model of TRPV1-RTX/DkTx complex (toxin is shown in magenta and indicated by arrows) built as described in Methods.



Extended Data Figure 6 | 3D reconstruction of TRPV1-capsaicin complex low-pass filtered at 4.2 Å resolution. a–d, Four different views of the 3D reconstruction low-pass filtered to 4.2 Å with a temperature factor of -150 Å^2 ,

fitted with *de novo* atomic model of TRPV1-capsaicin complex built as described in Methods.

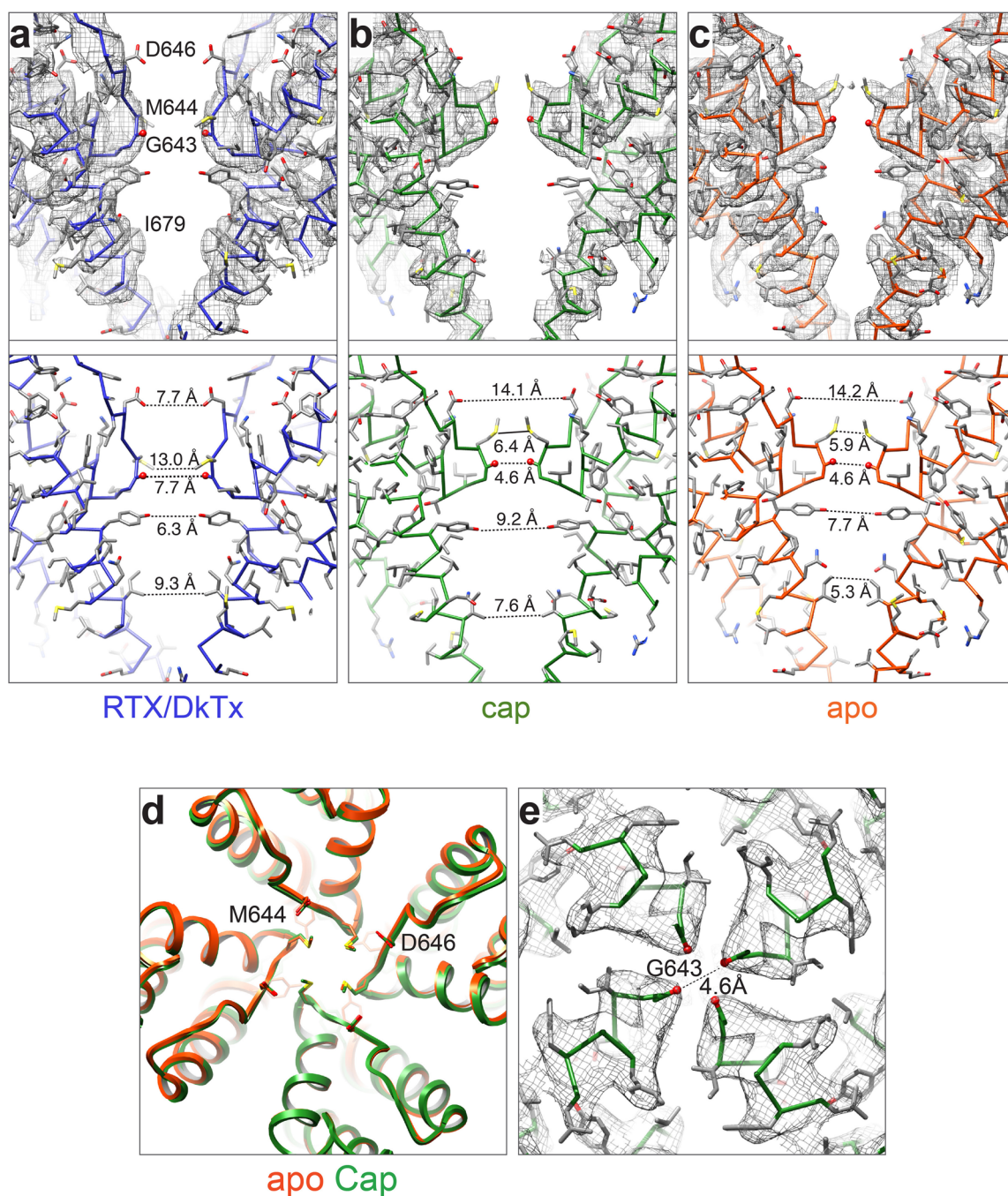


Extended Data Figure 7 | Observed densities in vanilloid pocket.

a, Non-protein associated densities in the region adjacent to S3–S4 transmembrane helices observed in 3D density maps of the apo TRPV1 structure (blue, 3.4 Å, -100 Å^2) or TRPV1 in complex with RTX/DkTx (red, 3.8 Å, -150 Å^2) or capsaicin (yellow, 3.9 Å, -150 Å^2), as indicated.

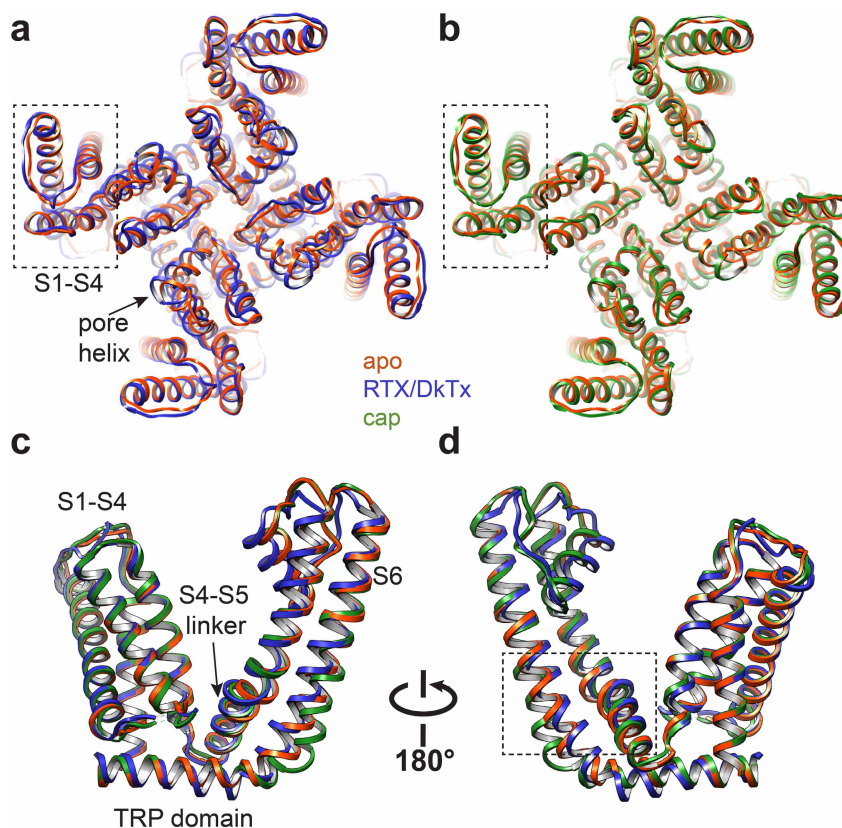
b, Density of bound capsaicin (blue) viewed from the side (left) and top-down

(that is, from the extracellular face; right). Density is also observed in the apo-channel structure (purple), possibly representing an endogenous lipid or other small hydrophobic molecule. All maps were low-pass filtered to 4.5 Å with a temperature factor of -200 Å^2 , normalized and displayed at the same sigma level (8σ).



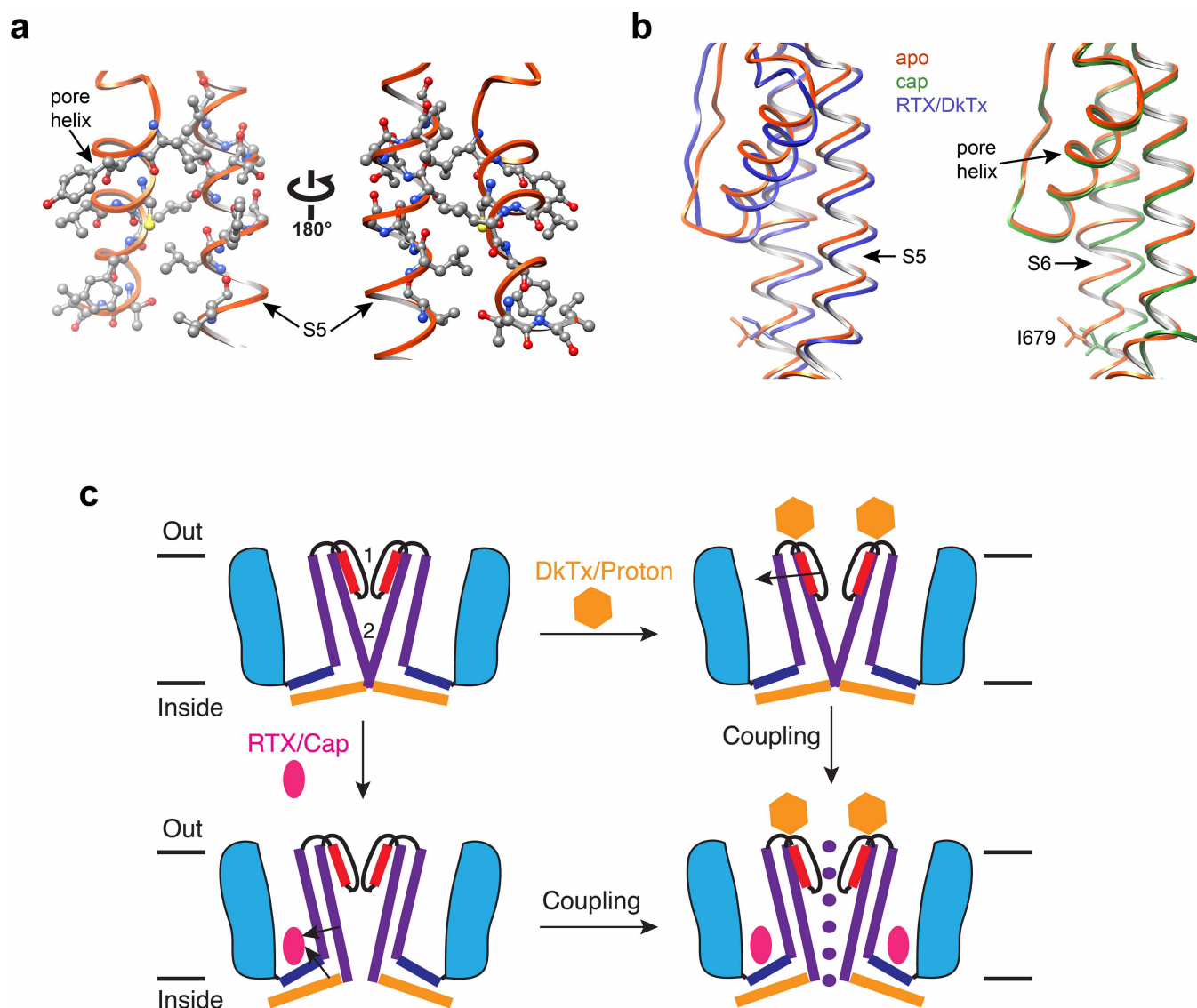
Extended Data Figure 8 | Structural details of the TRPV1 pore with and without ligands. **a–c**, Density maps for pore region for two diagonally opposed monomers superimposed onto their atomic models (top). Distances between specific side-chain atoms along the pore are also indicated (bottom). **d**, Superimposed top-down view of apo and capsaicin-bound TRPV1 outer

pore regions (orange and green, respectively). **e**, Density map of selectivity filter in the capsaicin-bound TRPV1 structure. The distance between carbonyl oxygens of diagonally opposed G643 residues (4.6 Å) does not differ from that of the apo structure (4.6 Å).



Extended Data Figure 9 | S1-S4 as a stationary domain. **a**, Superimposition of apo and RTX/DkTx-bound TRPV1 structures (orange and blue, respectively) from top-down view. S1-S4 domain (outlined in dashed box) shows near-complete overlap. **b**, Same comparison for apo and

capsaicin-bound channel structures (orange and green, respectively). **c**, **d**, Superimposition of transmembrane core of apo versus RTX/DkTx- or capsaicin-bound TRPV1 structures (orange, blue and green, respectively). Dashed box denotes region highlighted in Fig. 6.



Extended Data Figure 10 | Dual gate model for TRPV1 activation. **a**, Pore helix and upper half of S5 helix are in close proximity and appear to be physically coupled, representing a potential mechanism for allosteric coupling between upper and lower gates. Several residues on both helices are rendered as ball-and-stick to highlight close apposition. **b**, Downward tilt of pore helix away from the central pore is associated with movement of the S5 helix in RTX/DkTx structure (left). This structural arrangement is not observed in capsaicin-bound structure (right). **c**, Model depicting two gate mechanism of TRPV1 activation. Two main constriction points at the selectivity filter (1) and lower gate (2) block ion permeation in the apo, closed state (top left). Some pharmacological agents (for example, protons or spider toxins; gold hexagon)

target the outer pore region of the channel to open or stabilize the conductive conformation of the selectivity filter (top right). Arrow denotes proposed coupling between the pore helix and S5. Small vanilloid ligands (for example, RTX and capsaicin; red ellipse) bind within a hydrophobic pocket formed by the S3–S4 helices, the S4–S5 linker and the pore module, eliciting conformational changes that expand the lower gate (bottom left). Arrows indicate proposed coupling between S4–S5 helix, S6 and TRP domain. Full expansion of the ion permeation pathway and ion conduction is achieved when both upper and lower gates are opened (bottom right). Pharmacological and mutagenesis data suggest that these gates are allosterically coupled.

Highly polarized light from stable ordered magnetic fields in GRB 120308A

C. G. Mundell¹, D. Kopač², D. M. Arnold¹, I. A. Steele¹, A. Gomboc^{2,3}, S. Kobayashi¹, R. M. Harrison¹, R. J. Smith¹, C. Guidorzi⁴, F. J. Virgili¹, A. Melandri⁵ & J. Japelj²

After the initial burst of γ -rays that defines a γ -ray burst (GRB), expanding ejecta collide with the circumburst medium and begin to decelerate at the onset of the afterglow, during which a forward shock travels outwards and a reverse shock propagates backwards into the oncoming collimated flow, or ‘jet’^{1,2}. Light from the reverse shock should be highly polarized if the jet’s magnetic field is globally ordered and advected from the central engine^{3,4}, with a position angle that is predicted to remain stable in magnetized baryonic jet models⁵ or vary randomly with time if the field is produced locally by plasma or magnetohydrodynamic instabilities^{6,7}. Degrees of linear polarization of $P \approx 10$ per cent in the optical band have previously been detected in the early afterglow^{6,8}, but the lack of temporal measurements prevented definitive tests of competing jet models^{9–14}. Hours to days after the γ -ray burst, polarization levels are low ($P < 4$ per cent), when emission from the shocked ambient medium dominates^{15–17}. Here we report the detection of $P = 28^{+4}_{-4}$ per cent in the immediate afterglow of Swift γ -ray burst GRB 120308A, four minutes after its discovery in the γ -ray band, decreasing to $P = 16^{+5}_{-4}$ per cent over the subsequent ten minutes. The polarization position angle remains stable, changing by no more than 15 degrees over this time, with a possible trend suggesting gradual rotation and ruling out plasma or magnetohydrodynamic instabilities. Instead, the polarization properties show that GRBs contain magnetized baryonic jets with large-scale uniform fields that can survive long after the initial explosion.

On 8 March 2012 at $T_0 = 06:13:38$ UT, NASA’s Swift satellite identified GRB 120308A as a single, broad pulse of γ -rays lasting approximately 100 s, beginning at $T_0 - 30$ s, peaking at $\sim T_0 + 1$ s and ending at $\sim T_0 + 70$ s (ref. 18). The Swift X-ray telescope began observing the GRB at $T_0 + 92.6$ s, identifying a bright X-ray afterglow with a light curve exhibiting the canonical behaviour of a typical long GRB. The Liverpool Telescope responded automatically to the Swift trigger and identified the optical afterglow¹⁹.

Polarimetry was performed using the purpose-built RINGO2 polarimeter²⁰ on the Liverpool Telescope (see Supplementary Information). RINGO2 observations of GRB 120308A started at 06:17:38 UT ($T_0 + 240$ s) and ended at 06:27:25 UT. During that time, a total of 5,600 images were taken (700 at each angle of the Polaroid polarizer; Fig. 1a). After correcting for instrumental effects, co-adding the data at each rotator angle over that period showed a strong time-averaged polarization signal from the GRB of $P \approx 20\%$ compared with values $P < 3\%$ for the other objects of similar brightness in the image (Fig. 1b).

The time-sampled polarization over this period is both high and variable, with $P = 28 \pm 4\%$ declining to $P = 16^{+5}_{-4}\%$ by ~ 800 s after the GRB trigger (Fig. 2a), in contrast to unpolarized comparison objects in the GRB field of view, which straddle the GRB brightness and location and show no significant variation over the same time. The GRB polarization position angle θ is remarkably stable over this period (Fig. 2b), with a total variation that does not exceed $\theta \approx 15^\circ$ and which shows a

trend that may be consistent with gradual rotation of the angle, ruling out plasma or magnetohydrodynamic instabilities^{6,7}. The measured polarization evolution is also robust to different choices of temporal binning. The derived extinction local to the GRB would induce less than 1% of polarization in the GRB afterglow; the high detected polarization in GRB 120308A and the observed temporal variation are therefore intrinsic to the GRB and not due to dust scattering or instrumental effects (see Supplementary Information).

Following the RINGO2 exposures, a series of Liverpool Telescope RATCam optical images was collected until twilight made data collection no longer possible. The corresponding time evolution of the optical flux density using measurements from both RINGO2 and RATCam is shown as a light curve in Fig. 2c, in which the optical flux density reaches a peak at $\sim T_0 + 300$ s, followed by a steady fade of the emission with a possible plateau around $\sim T_0 + 1,000$ s.

In the standard ‘reverse plus forward shock’ scenario², emission from both shocks contribute to the afterglow; their relative brightnesses and temporal evolution combine to produce the observed light curve. Empirically, optical light curves are commonly fitted with a collection of double broken power laws²¹ in an attempt to parametrize the individual contributions from reverse and forward shocks to the overall shape of the light curve. The light curve of GRB 120308A can either be parametrized with a single peak (which contains equal contributions from reverse and forward shock emission²²) or with two peaks (of different brightness and separated in time⁸ such that the reverse shock peak dominates the light curve at early time and the later emergence of the forward shock accounts for the flattening of the light curve at $\sim 1,000$ s; see Supplementary Information). In both cases, the peak at $\sim T_0 + 300$ s represents the onset of deceleration of the fireball and the emission at that time contains a significant contribution from the reverse shock.

Using the relative strengths of reverse- and forward-shock emission in both scenarios, we derive an independent estimate of the jet’s magnetization²³; the magnetic energy density in the reverse-shock region is higher than in the forward-shock region by a factor of ~ 30 for the one-peak model and by a factor > 500 for the two-peak case. In both cases, a magnetized reverse shock is required, consistent with the high degree of polarization detected.

Figure 3 shows degree of polarization as a function of time after the burst in the cosmological rest frame; GRB 120308A is shown in the context of previous polarization measurements of other GRBs at early and late times. At $P = 28\%$, the optical polarization measured in GRB 120308A is significantly higher than previous detections. Tangled magnetic fields produced locally in a shock front produce low net polarization of a few per cent (ref. 15), as seen at late time, and random temporal change in polarization angle. They are therefore excluded as the origin of polarized optical emission in GRB 120308A. Instead, a magnetized reverse shock with an ordered magnetic field is needed to explain the large net polarization and stable position angle in GRB 120308A.

¹Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK. ²Faculty of Mathematics and Physics, University of Ljubljana, Jadranska ulica 19, 1000 Ljubljana, Slovenia. ³Centre of Excellence SPACE-SI, Aškerčeva cesta 12, 1000 Ljubljana, Slovenia. ⁴Physics Department, University of Ferrara, Via Saragat, 1, 44122 Ferrara, Italy. ⁵INAF/Brera Astronomical Observatory, via Biancamano 46, 23807, Merate (LC), Italy.

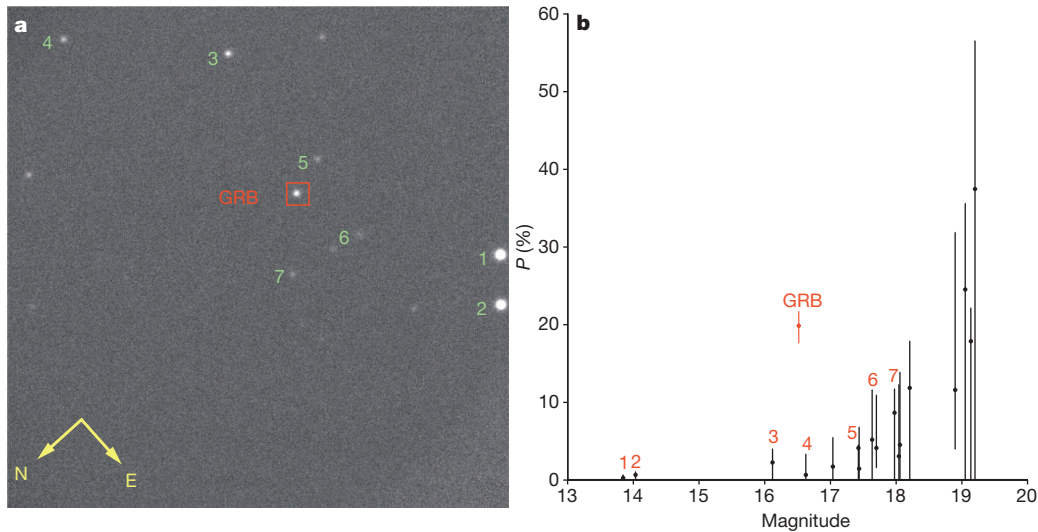


Figure 1 | Time-integrated optical properties of the GRB 120308A field. **a**, RINGO2 total intensity image of $4' \times 4'$ field containing GRB 120308A, with total exposure time 588 s. The GRB (boxed) and seven comparison objects (numbered) are indicated; the directions of north and east are shown. RINGO2 combines a Polaroid polarizer rotating at ~ 1 revolution s^{-1} with a fast readout electron multiplying CCD camera that is triggered eight times per revolution. Summing data from each rotation angle allows derivation of the total intensity for each source in the image, while analysis of their relative intensities allows calculation of their Stokes parameters²⁵. Measurements are not affected by variations in source brightness or observing conditions on timescales > 1 s owing to the rapid rotation of the polaroid. There is no significant variation in atmospheric transparency or seeing (image point-spread function) over the 588-s exposure. **b**, Measured time-averaged polarization P of all objects versus

apparent magnitude. As P is a one-sided (always positive) quantity, noise in the Stokes q and u parameters translates into a rising P with large uncertainty for the faintest objects, even though their actual polarization is likely to be small. The strong time-averaged polarization of the GRB (red symbol) of 20% compared to sources of similar brightness is obvious. Error bars ($\pm 1\sigma$) were calculated using a Monte Carlo simulation ($N = 10,000$). This used a range of input q and u values with an error distribution calculated from the combination of photon counting statistics with the uncertainty in instrumental calibration to calculate 1σ ranges of P and position angle (θ) for each object. All quoted measurements in this Letter use this Monte Carlo estimator, although because polarization in GRB 120308A is significantly non-zero, the derived errors (within $\sim 1\%$ absolute error) are comparable to standard error analyses for that object (see Supplementary Information and Extended Data Figs 1, 2, 3, 4, 5, 6).

The theoretical maximum degree of linear polarization of synchrotron radiation emitted by electrons in a perfectly homogeneous magnetic field is $P \approx 70\%$; the difference between the measured and the theoretical maximum can therefore provide further constraints on the

physical properties of the emitting source. The measured net polarization can be less than the theoretical maximum because of (1) the dilution of polarized reverse-shock emission by unpolarized forward-shock emission, (2) the combination of ordered magnetic fields from the central

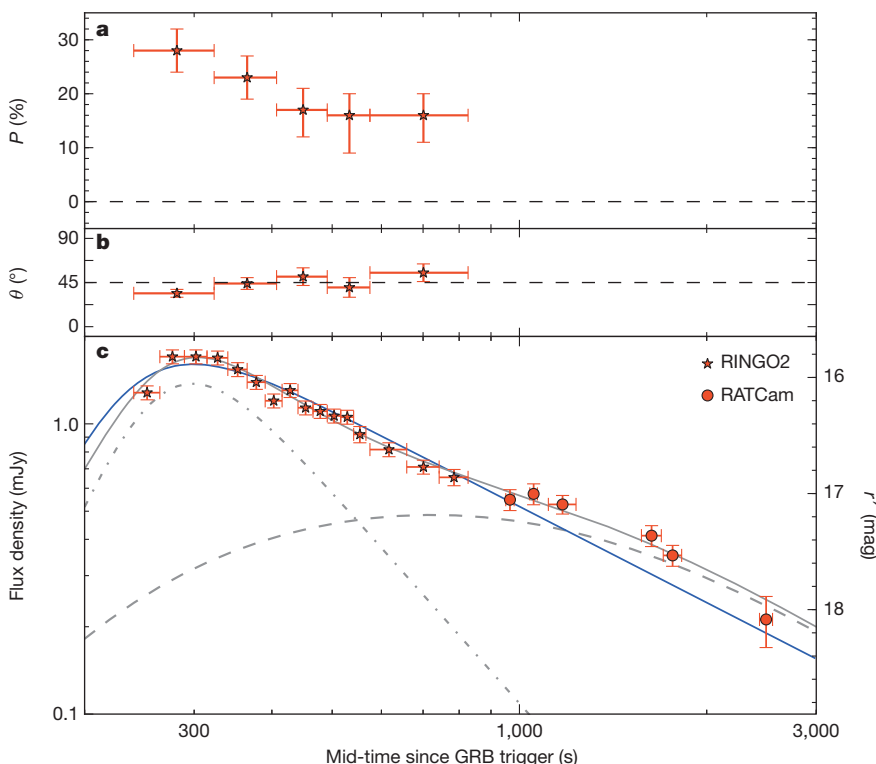


Figure 2 | Evolution of optical polarization and brightness in GRB 120308A. **a**, **b**, Evolution of polarization degree P (**a**) and position angle θ (**b**; degrees east of north) for GRB 120308A. Individual 0.125-s RINGO2 exposures at the eight Polaroid angles are co-added over a desired time interval into eight images, on which absolute aperture photometry is performed and P and θ derived. Owing to the low read noise of the system, data from each rotation angle can be stacked into temporal bins after data acquisition to optimize signal-to-noise ratio versus time resolution. Here the data were subdivided into four bins of duration ~ 84 s and one bin of ~ 252 s giving roughly equal signal-to-noise ratio. The observed polarization properties are robust to alternative choices of temporal binning (see Supplementary Information and Extended Data Figs 7, 8, 9). Error bars, $\pm 1\sigma$, as described in Fig. 1b. **c**, Light curve of GRB 120308A in red (555–690 nm) light using RINGO2 and RATCam. Data have been cross-calibrated to the SDSS r' system via five objects in common, with a possible systematic error of up to $\sim 6\%$ between the two instruments due to colour effects. Model fits using one peak (blue solid line) or two peaks (broken grey line for each component; resultant combined light curve in solid grey) are shown with an additional point²⁶ constraining late time behaviour (see Supplementary Information). The two-peak model is statistically slightly preferred. Error bars, $\pm 1\sigma$.

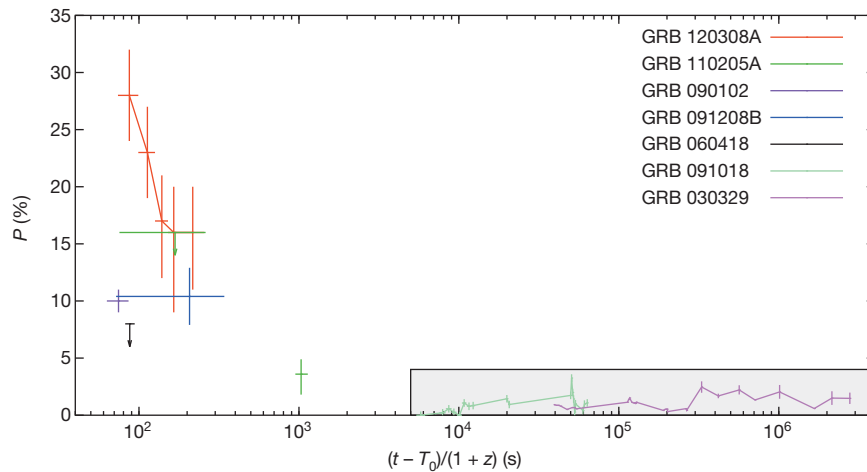


Figure 3 | Rest-frame optical polarization properties of GRBs. The degree of optical polarization P is plotted as a function of time after the burst in the cosmological rest frame, $(t - T_0)/(1 + z)$, where z is redshift and $(t - T_0)$ is the time after the burst in the observer frame. GRB 120308A, GRB 110205A²⁷, GRB 091208B⁶, GRB 090102⁸ and GRB 060418²² were measured at early time. The shaded area shows the typical polarization levels of GRBs measured at late times; representative examples GRB 091018¹⁷ and GRB 030329¹⁶ are shown. Polarization error bars are as reported in the corresponding publications; the temporal error bars show the duration of the measurement. See Supplementary Information and Extended Data Fig. 10 for the determination of redshift for GRB 120308A.

engine and tangled local magnetic fields, (3) a toroidal magnetic field viewed slightly off-axis to the jet axis or (4) large-scale magnetic fields (including toroidal field) that are distorted on an angular scale $1/\Gamma$ corresponding to the relativistically beamed observable scale around the line of sight (here Γ is the Lorentz factor). All four scenarios could apply to the single-peak model. In the two-peak model, scenarios (1) and (2) are excluded owing to the dominant reverse-shock emission and high magnetization. Regardless of the particular model, the polarization characteristics reported here probe an important phase—previously unseen—in the evolution of the physical conditions in the relativistic jet. The large-scale field required in GRB 120308A may not be perfectly homogeneous, even in the small observable angular scale $\sim 1/\Gamma$, and small distortions in the field lines could cause a slight change in the polarization position angle; the trend observed in GRB 120308A will provide new constraints on jet models.

We note that although the radiation mechanism responsible for prompt γ -ray emission has not yet been firmly established, the large values of γ -ray polarization in GRB 110301A and GRB 110721A²⁴, if confirmed in other GRBs, provide complementary evidence for magnetic fields in the ejecta. The high degree of optical polarization detected in GRB 120308A and its stable position angle shows that large-scale fields survive long after the burst. In the future, detection of optical, γ -ray and microwave polarization properties in the same GRB would provide valuable insight into the full evolution of the magnetic field.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 July; accepted 25 October 2013.

- Piran, T. Gamma ray bursts and the fireball model. *Phys. Rep.* **314**, 575–667 (1999).
- Zhang, B., Kobayashi, S. & Mészáros, P. Gamma-ray burst early optical afterglows: implications for the initial Lorentz factor and the central engine. *Astrophys. J.* **595**, 950–954 (2003).
- Granot, J. & Königl, A. Simulations of ultrarelativistic magnetodynamic jets from gamma-ray burst engines. *Astrophys. J.* **594**, L83–L87 (2003).
- Lytikov, M. Explosive reconnection in magnetars. *Mon. Not. R. Astron. Soc.* **346**, 540–554 (2003).
- Lazzati, D. *et al.* On the jet structure and magnetic field configuration of GRB 020813. *Astron. Astrophys.* **422**, 121–128 (2004).
- Uehara, T. *et al.* GRB 091208B: first detection of the optical polarization in early forward shock emission of a gamma-ray burst afterglow. *Astrophys. J.* **752**, L6 (2012).
- Gruzinov, A. & Waxman, E. Gamma-ray burst afterglow: polarization and analytic light curves. *Astrophys. J.* **511**, 852–861 (1999).
- Steele, I. A., Mundell, C. G., Smith, R. J., Kobayashi, S. & Guidorzi, C. Ten percent polarized optical emission from GRB 090102. *Nature* **462**, 767–769 (2009).
- Medvedev, M. V. & Loeb, A. Generation of magnetic fields in the relativistic shock of gamma-ray burst sources. *Astrophys. J.* **526**, 697–706 (1999).
- Lytikov, M. The electromagnetic model of gamma-ray bursts. *New J. Phys.* **8**, 119–123 (2006).

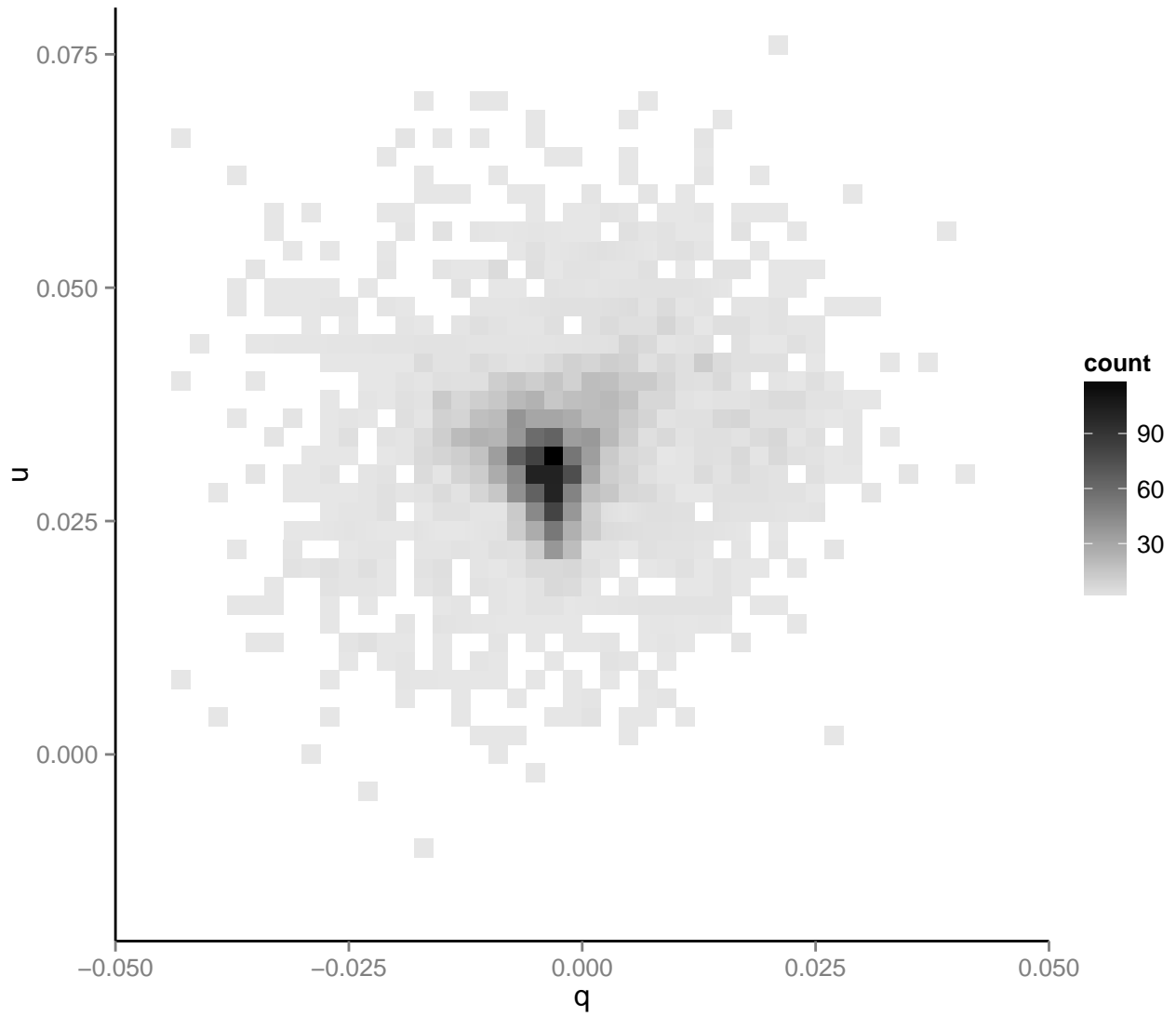
- Granot, J. The effects of sub-shells in highly magnetized relativistic flows. *Mon. Not. R. Astron. Soc.* **421**, 2467–2477 (2012).
- Komissarov, S. S. Shock dissipation in magnetically dominated impulsive flows. *Mon. Not. R. Astron. Soc.* **422**, 326–346 (2012).
- Zhang, B. & Yan, H. The internal-collision-induced magnetic reconnection and turbulence (ICMART) model of gamma-ray bursts. *Astrophys. J.* **726**, 90 (2011).
- Tchekhovskoy, A., McKinney, J. C. & Narayan, R. Simulations of ultrarelativistic magnetodynamic jets from gamma-ray burst engines. *Mon. Not. R. Astron. Soc.* **388**, 551–572 (2008).
- Covino, S. GRB 990510: linearly polarized radiation from a fireball. *Astron. Astrophys.* **348**, L1–L4 (1999).
- Greiner, J. *et al.* Evolution of the polarization of the optical afterglow of the γ -ray burst GRB 030329. *Nature* **426**, 157–159 (2003).
- Wiersema, K. *et al.* Detailed optical and near-infrared polarimetry, spectroscopy and broad-band photometry of the afterglow of GRB 091018: polarization evolution. *Mon. Not. R. Astron. Soc.* **426**, 2–22 (2012).
- Baumgartner, W. H. *et al.* GRB 120308A: Swift detection of a burst. *GCN Circ.* **13017** (2012).
- Virgili, F. *et al.* GRB 120308A: Liverpool Telescope optical afterglow candidate. *GCN Circ.* **13018** (2012).
- Steele, I. A. *et al.* RINGO2: an EMCCD-based polarimeter for GRB followup. *Proc. SPIE* **7735**, 773549 (2010).
- Beuermann, K. VLT observations of GRB 990510 and its environment. *Astron. Astrophys.* **352**, L26–L30 (1999).
- Mundell, C. G. *et al.* Early optical polarization of a gamma ray burst afterglow. *Science* **315**, 1822–1824 (2007).
- Harrison, R. M. & Kobayashi, S. Magnetization degree of gamma-ray burst fireballs: numerical study. *Astrophys. J.* **772**, 101 (2013).
- Yonetoku, D. *et al.* Magnetic structures in gamma-ray burst jets probed by gamma-ray polarization. *Astrophys. J.* **758**, L1 (2012).
- Clarke, D. & Neumayer, D. Experiments with a novel CCD stellar polarimeter. *Astron. Astrophys.* **383**, 360–366 (2002).
- Bikmaev, I. *et al.* GRB 120308A: RTT150 optical observations. *GCN Circ.* **13030** (2012).
- Cucchiara, A. *et al.* Constraining gamma-ray burst emission physics with extensive early-time, multiband follow-up. *Astrophys. J.* **743**, 154 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements The Liverpool Telescope is operated by Liverpool John Moores University at the Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. C.G.M. acknowledges support from the Royal Society, the Wolfson Foundation and the Science and Technology Facilities Council. A.G. acknowledges funding from the Slovenian Research Agency and from the Centre of Excellence for Space Sciences and Technologies SPACE-SI, an operation partly financed by the European Union, European Regional Development Fund and Republic of Slovenia, Ministry of Education, Science and Sport.

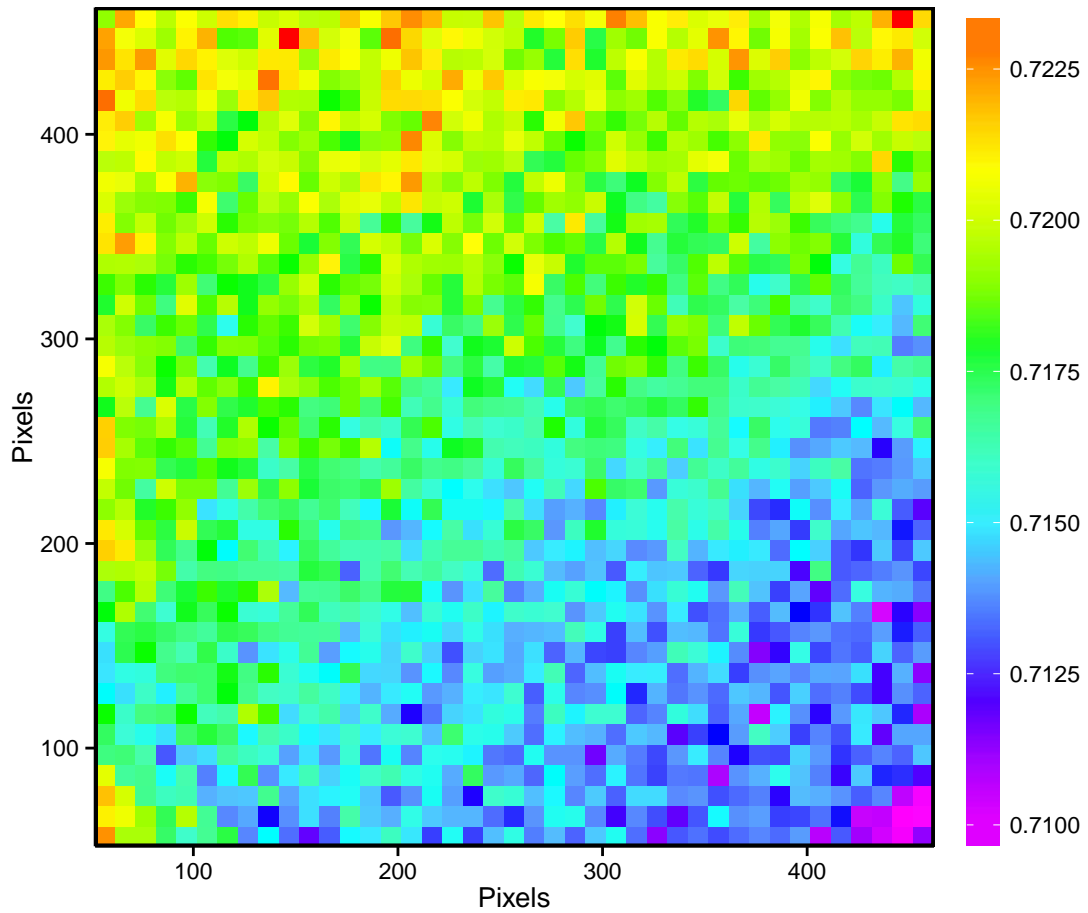
Author Contributions C.G.M.: instrument science case, observations, scientific interpretation, paper writing lead. I.A.S.: instrument design and build, data calibration. D.K., D.M.A., R.J.S., A.M.: data reduction, analysis and instrument calibration. A.G.: science case, observations, scientific interpretation. C.G., F.J.V., J.J.: observations, data analysis, energy, redshift derivations and afterglow identification. S.K., R.M.H.: scientific and theoretical interpretation. All authors contributed to the writing/editing of the paper and overall scientific interpretation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.G.M. (c.mundell@ljmu.ac.uk).

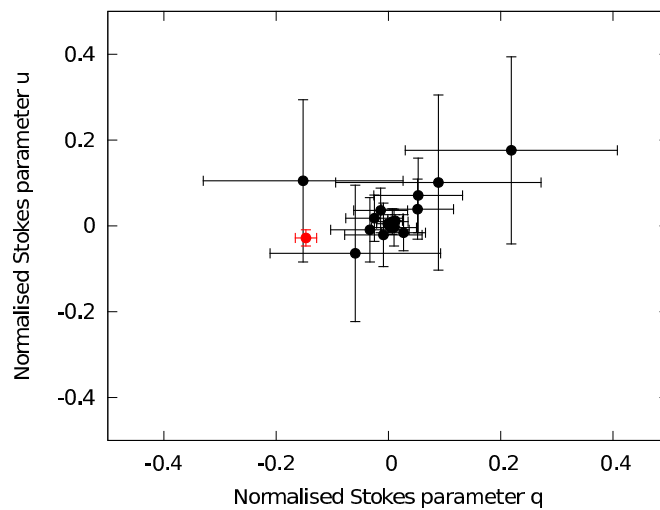


Extended Data Figure 1 | RINGO2 measurements of Stokes parameters for zero-polarized stars. Shown is a two-dimensional histogram depicting the distribution of q values versus u values, with bin size 0.002 in both coordinates,

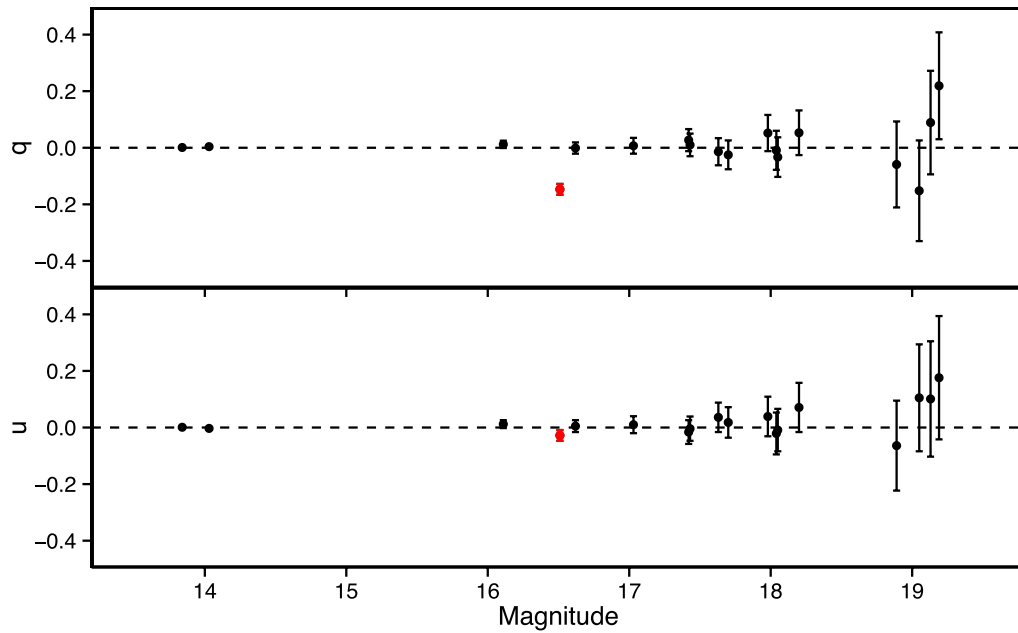
for 3,955 objects with $r' < 16$ mag detected in $\sim 2,000$ observations of zero-polarized standard star fields. See Supplementary Information for more details.



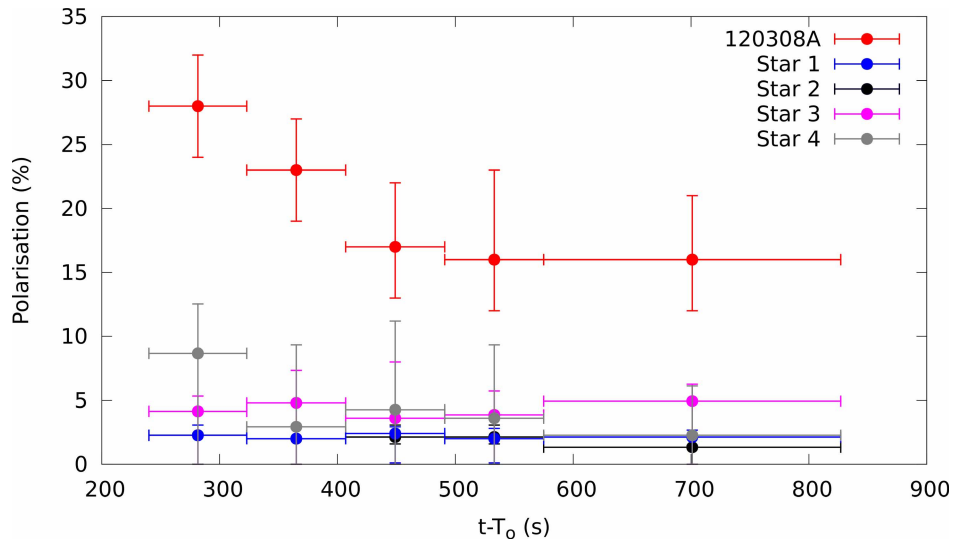
Extended Data Figure 2 | RINGO2 measurement of the polarization of the zenith sky at sunset over a $4' \times 4'$ field of view. See Supplementary Information for more details.



Extended Data Figure 3 | RINGO2 measurement of Stokes parameters for all objects in the GRB field. The complete data set (all temporal bins) has been used. The GRB optical counterpart is indicated in red. The four points with large error bars are too faint ($r' > 18.5$ mag) for reliable measurements to be made. Error bars are given at 1σ confidence. See Supplementary Information for more details.

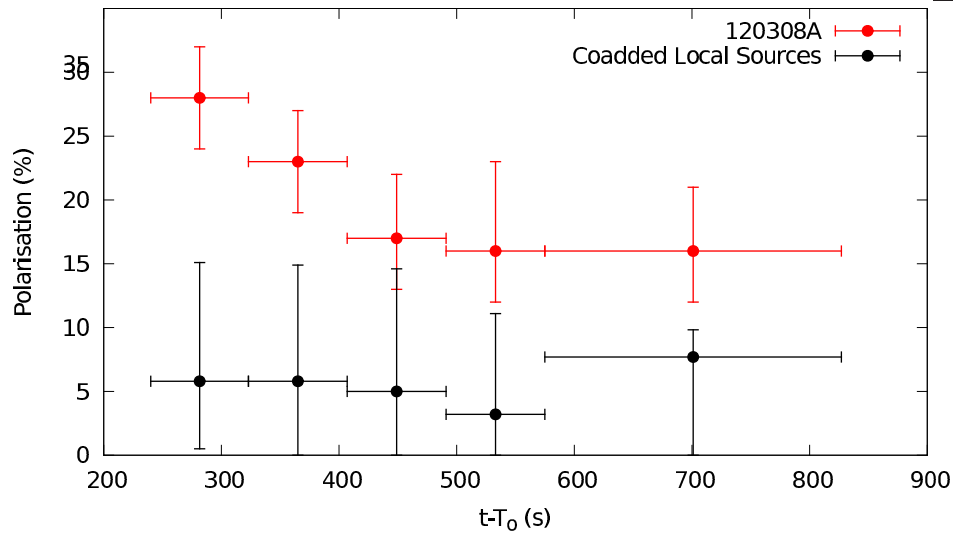


Extended Data Figure 4 | Stokes q and u parameters of all objects in the GRB field versus r' magnitude. Error bars are given at 1σ confidence. See Supplementary Information for more details.



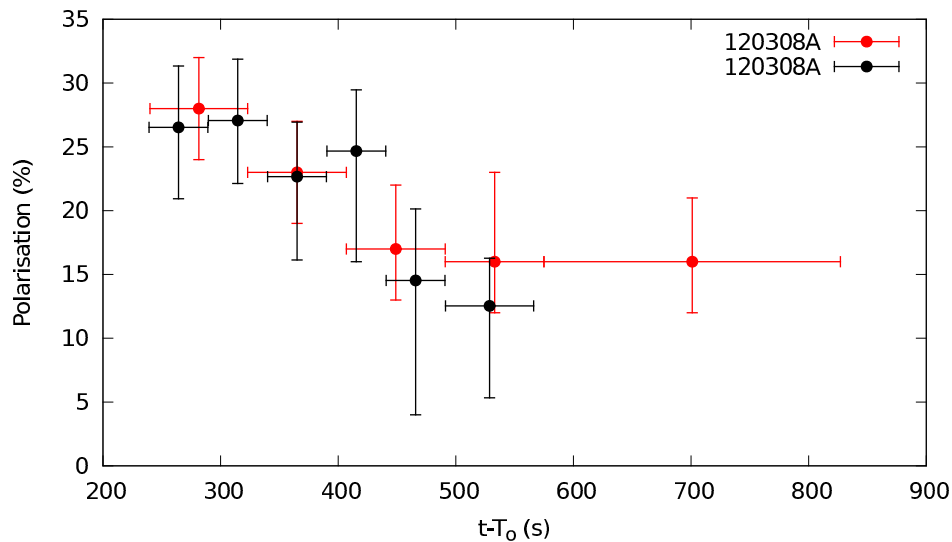
Extended Data Figure 5 | Polarization evolution of GRB 120308A compared with the four brightest objects 1–4 (Fig. 1) in the field. The data have been split into the same five temporal bins as presented in the main text.

In comparison with the GRB, no significant variation of the comparison objects is apparent. Error bars are given at 1σ confidence. See Supplementary Information for more details.



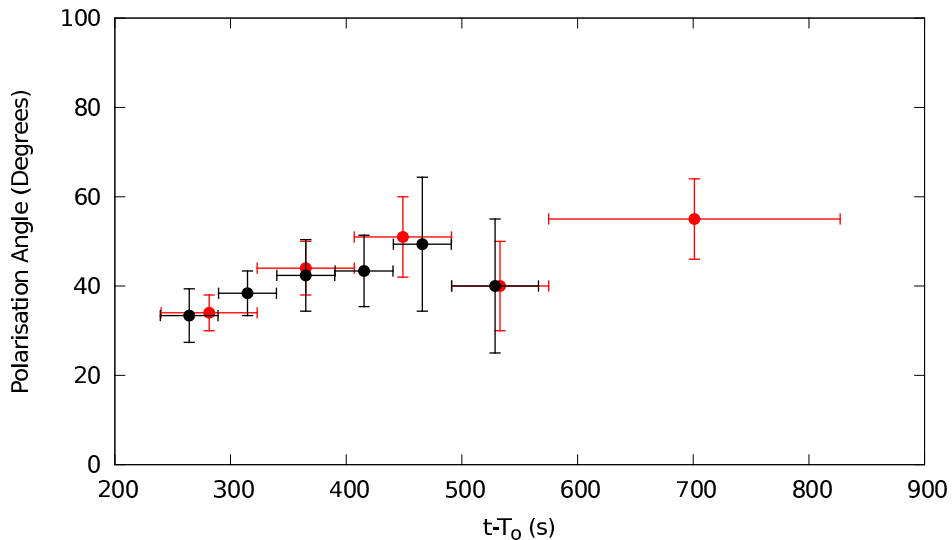
Extended Data Figure 6 | Polarization evolution of GRB 120308A compared with a co-addition of three nearby objects in the field. In order to enhance the signal-to-noise ratio, the data for objects 5–7 (Fig. 1) have been

co-added to produce a source of apparent magnitude similar to that of the GRB. The data have been split into the same five temporal bins as presented in the main text. See Supplementary Information for more details.



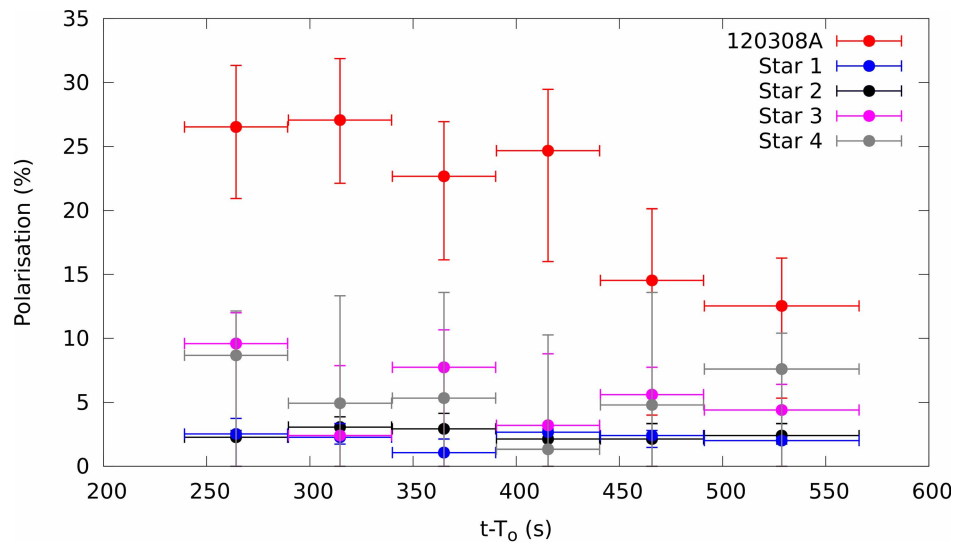
Extended Data Figure 7 | Polarization degree evolution of GRB 120308A with two different temporal binning schemes. Shown is original (red) and an alternative co-addition (black) of the data over the time period covered by

the first four temporal bins into six bins. The polarization evolution of the optical counterpart is unaffected within the error bars. Error bars are given at 1σ confidence. See Supplementary Information for more details.



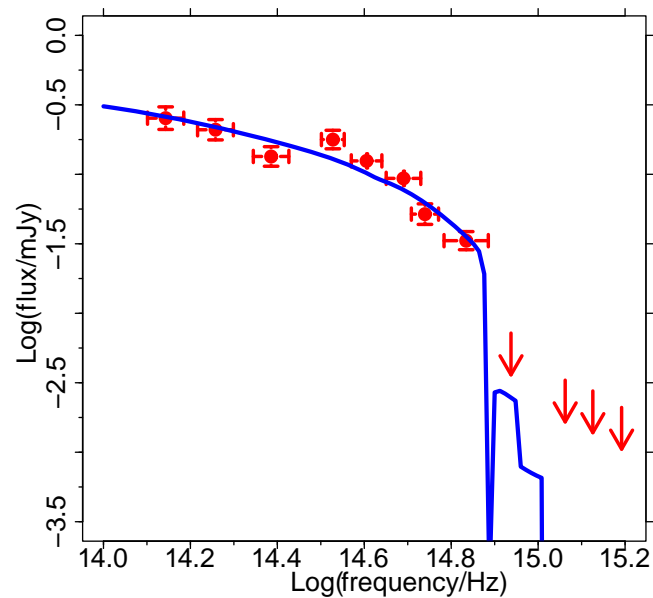
Extended Data Figure 8 | Polarization angle evolution of GRB 120308A with two different temporal binning schemes. Original (red) and an alternative co-addition (black) of the data over the time period covered by the

first four temporal bins into six bins. The evolution of the polarization angle is unaffected within the error bars. Error bars are given at 1σ confidence. See Supplementary Information for more details.



Extended Data Figure 9 | Polarization evolution of GRB 120308A compared with the four brightest objects in the field for alternative temporal binning. The data have been split into six temporal bins covering the

first 305 s of observation. As in Extended Data Fig. 5, in comparison with the GRB no significant variation of the comparison objects is apparent. Error bars are given at 1σ confidence. See Supplementary Information for more details.



Extended Data Figure 10 | Spectral energy distribution and best fit model for GRB 120308A. Flux density at $t = 1.44$ h after the GRB is plotted as a function of observing frequency. Upper limits are given at 3σ confidence. Error bars are given at 1σ confidence. See Supplementary Information for more details.

Olivine in an unexpected location on Vesta's surface

E. Ammannito¹, M. C. De Sanctis¹, E. Palomba¹, A. Longobardo¹, D. W. Mittlefehldt², H. Y. McSween³, S. Marchi^{1,4}, M. T. Capria¹, F. Capaccioni¹, A. Frigeri¹, C. M. Pieters⁵, O. Ruesch⁶, F. Tosi¹, F. Zamboni¹, F. Carraro¹, S. Fonte¹, H. Hiesinger⁶, G. Magni¹, L. A. McFadden⁷, C. A. Raymond⁸, C. T. Russell⁹ & J. M. Sunshine¹⁰

Olivine is a major component of the mantle of differentiated bodies, including Earth. Howardite, eucrite and diogenite (HED) meteorites represent regolith, basaltic-crust, lower-crust and possibly ultramafic-mantle samples of asteroid Vesta, which is the lone surviving, large, differentiated, basaltic rocky protoplanet in the Solar System¹. Only a few of these meteorites, the orthopyroxene-rich diogenites, contain olivine, typically with a concentration of less than 25 per cent by volume². Olivine was tentatively identified on Vesta^{3,4}, on the basis of spectral and colour data, but other observations did not confirm its presence⁵. Here we report that olivine is indeed present locally on Vesta's surface but that, unexpectedly, it has not been found within the deep, south-pole basins, which are thought to be excavated mantle rocks^{6–8}. Instead, it occurs as near-surface materials in the northern hemisphere. Unlike the meteorites, the olivine-rich (more than 50 per cent by volume) material is not associated with diogenite but seems to be mixed with howardite, the most common^{7,9} surface material. Olivine is exposed in crater walls and in ejecta scattered diffusely over a broad area. The size of the olivine exposures and the absence of associated diogenite favour a mantle source, but the exposures are located far from the deep impact basins. The amount and distribution of observed olivine-rich material suggest a complex evolutionary history for Vesta.

The Visible and Infrared Mapping Spectrometer (VIR) on board NASA's Dawn spacecraft¹⁰ has been used in a global search for olivine on the Vestan surface (Supplementary Information). VIR revealed a global-scale dichotomy^{7,8} (Fig. 1), with diogenite-rich material exposed predominantly in the deeply excavated southern hemisphere. Magma-ocean models for Vesta's differentiation yield eucritic crust overlying a diogenite layer, with olivine-rich mantle rocks and a metallic core in the deep interior^{11,12}. These models predict mineralogical variations on a large vertical scale, with olivine-rich cumulates occurring below olivine-poor diogenite. Alternative models, more consistent with the diverse trace-element geochemistry of diogenites, posit that diogenitic plutons occur at the crust–mantle boundary or within the basaltic crust^{13,14}, resulting in association of olivine-rich and orthopyroxene-rich diogenites mixed on smaller scales.

VIR spectra did not provide definitive evidence for olivine within the two large basins in the southern hemisphere^{6–8}. However, typical olivine-bearing diogenites cannot be easily distinguished spectrally from olivine-free diogenites¹ because of the difficulty of identifying olivine at low concentrations in the presence of abundant orthopyroxene^{15,16}; thus, olivine may be present within the southern basins but only in modest amounts (≤ 25 vol%, comparable to that reported for most olivine-bearing diogenites²).

Unexpectedly, olivine-rich areas have now been discovered in the northern hemisphere. The VIR spectra of ejecta surrounding Arruntia crater and the nearby Bellicia crater (Fig. 2) reveal clear olivine signatures (Fig. 3a), with the 1- μ m band (hereafter BI) centred at slightly longer wavelength than the average Vesta spectrum, and the centre of the 2- μ m band (hereafter BII) is unchanged. Laboratory data demonstrate

that pyroxene features dominate the spectra of olivine–pyroxene mixtures^{15,16}. Only olivine contents of ≥ 50 vol% produce a shift in the centre of BI^{15,16}, and the centre of BII remains unchanged with admixture of olivine (Fig. 3b). The three parameters we used to interpret olivine–pyroxene mixtures are the positions of the respective centres of BI and BII and variations in the band area ratio^{16–18} (BAR). In the BI–BII diagram (Fig. 3c), Bellicia and Arruntia data lie distinctly off the linear HED trend, with high values for BI centres that reveal the presence of olivine. Because the BII-centre position reflects the composition of pyroxene in olivine–pyroxene mixtures and, in the Bellicia–Arruntia area, lies between those of eucrites and diogenites, we have determined that the olivine in this area is associated with the mixed lithology, howardite.

This situation is distinct from the olivine occurrence in HED meteorites, where only very small amounts of olivine (≤ 3 vol%) occur in howardites¹⁹ (with the exception of the paired PCA 02 howardites, which nevertheless contain at most $\sim 7\%$ olivine²⁰). Olivine in HED meteorites occurs mainly in diogenites, which range from orthopyroxenite to harzburgite to dunite²¹ (Extended Data Fig. 1 and Extended Data Table 1). This observation is consistent with the interpretation that HED meteorites sample lithologies from Vesta's southern hemisphere that are associated with material ejected from the two large basins²².

The Vestan olivine-rich spectra and derived parameters are consistent with a mixture of 50–80-vol% olivine with pyroxene occurring over a broad area of hundred-kilometre size, encompassing both Bellicia crater and Arruntia crater. Olivine-rich material occurs as several high-albedo patches hundreds of metres across located high on the walls of Bellicia crater (Fig. 2b–d). Some of these patches have positive relief compared with the adjacent wall, suggesting more competent material (Fig. 2f). Several fresh small craters (with diameters of order 100 m) superposed on Bellicia ejecta also have high-albedo annuli with olivine spectral signatures (Fig. 2e). Olivine-rich material at Arruntia crater is most common in the ejecta blanket (Fig. 2g, h). The geological setting suggests that olivine-rich lithologies occur as a bright layer partly obscured by slump deposits and regolith mixing of the surface. Unlike its occurrence in HED meteorites, mainly as a small volume fraction in diogenites, here a lithology rich in olivine (≥ 50 vol%) in patches hundreds of metres in size is mixed with howarditic regolith.

The detected olivine-rich materials have characteristics at odds with pre-Dawn ideas about Vestan olivine: they are not associated with diogenites, they are located far from deeply excavated terrains in the southern hemisphere and they occur in large patches extending hundreds of metres.

Both exogenic and endogenic origins are possible. An exogenic origin seems unlikely, considering how uncommon xenocrystic (chondritic) olivine is in howardites²⁰ and the rarity of olivine-rich asteroids in the main belt²³ (Supplementary Information). Also, the large patches seem inconsistent with the fact that impactors are normally disaggregated. On the other hand, endogenic olivine is a component of Vesta, as

¹Istituto di Astrofisica e Planetologia Spaziali, INAF, 00133 Rome, Italy. ²NASA Johnson Space Center, Houston, Texas 77058, USA. ³Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA. ⁴NASA Lunar Science Institute, Boulder, Colorado 80302, USA. ⁵Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA. ⁶Institut für Planetologie, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany. ⁷NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ⁸Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. ⁹University of California, Los Angeles, California 90095, USA. ¹⁰Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA.

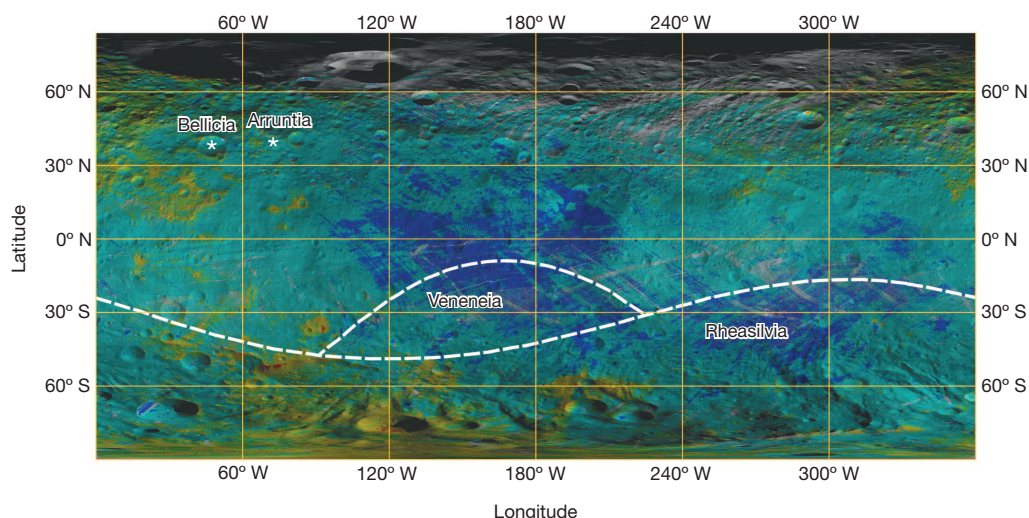


Figure 1 | HED meteorite distribution map. Lithological map of Vesta's surface derived from VIR spectra^{8,10} using all the data acquired during the Dawn orbital phases: red for diogenite, green for howardite, blue for eucrite, with overlapping fields of yellow for diogenitic howardite and cyan for eucritic howardite. The regions with magnesium-rich pyroxenes (red and yellow) correspond to a diogenite-dominated lithology. The distribution shows that the southern hemisphere is more rich in magnesian pyroxene with areas of nearly

demonstrated by its occurrence in diogenites and even in the PCA 02 howardites, where the target rock for olivine-bearing impact melts in these breccias was olivine-rich diogenite²⁰.

Two main models for the origin of endogenic olivine are serial-magmatism models that consider fractional crystallization in diogenite plutons emplaced at the base of, or within, the Vestan crust^{13,14}, and

pure diogenite, whereas the equatorial region and the northern hemisphere are more basalt-rich (eucritic). Howardites—brecciated mixtures of these lithologies—are the most abundant rocks observed on Vesta's surface. Arruntia and Bellicia craters are indicated, as well as the rim of Rheasilvia and Veneneia basins (dashed line). Howardites enriched in diogenites are visible in the ruined northern basins and in Rheasilvia (see Supplementary Information for further details and Extended Data Fig. 2).

magma-ocean models that predict an olivine-dominated mantle at depths of >20 – 40 km underlying an orthopyroxene-dominated (diogenitic) lower crust^{11,12}.

In the serial-magmatism hypothesis, a mixed region of eucrite, diogenite and olivine-rich material could have been sampled by impacts that did not excavate to great depth. The magma-ocean hypothesis

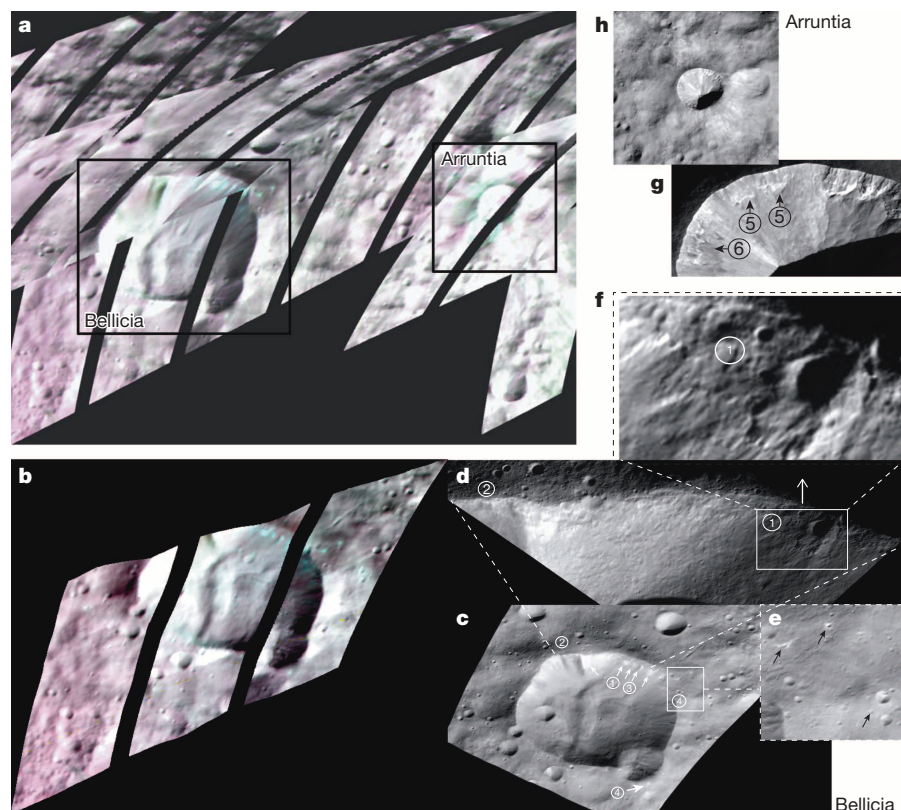
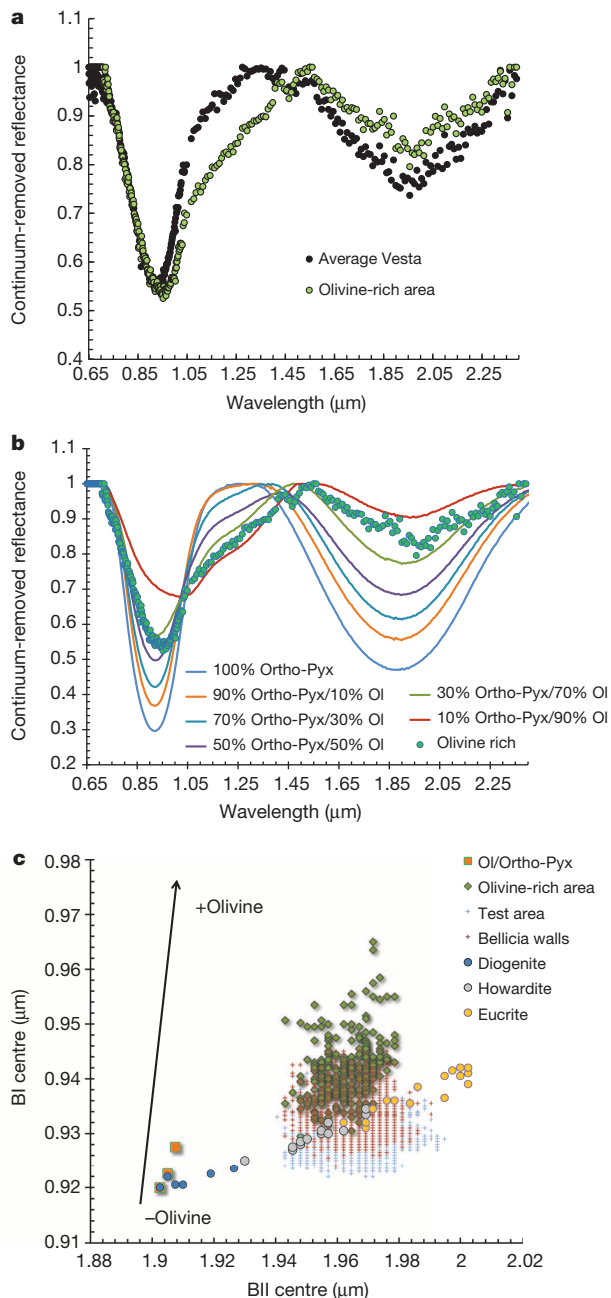


Figure 2 | Olivine-rich region in the visible and near-infrared wavelengths. **a**, Infrared mosaics of VIR data (longitudes 30° – 77° W, latitudes 25° – 60° N). Bellicia crater ($38\text{ km} \times 43\text{ km}$, longitude 48° W, latitude 38° N) and Arruntia crater ($\sim 11\text{-km}$ diameter, longitude 72° W, latitude 40° N) are enclosed in black squares. Coordinates in Claudia system. The mosaics have been made using VIR data from different observation cycles. The false colours (red, $1.25\text{ }\mu\text{m}$; green, $1.93\text{ }\mu\text{m}$; blue, $1.64\text{ }\mu\text{m}$), emphasize in green the olivine-rich region. **b**, Stretched view of Bellicia crater in false colours (same as in **a**) showing in green the purest olivine exposures. **c**–**f**, Framing Camera images of Bellicia crater. **c**, Image ($\sim 65\text{ m}$ per pixel) highlighting example locations of bright materials carrying the olivine-rich spectral signature (arrows). Numbered arrows show (1) olivine-rich material associated with a slump deposit downslope from a small crater; (2) comingling bright and dark materials on the crater wall; (3) relatively dark material adjacent to olivine-rich material; and (4) small craters on Bellicia ejecta that expose olivine-rich material. **d**, Image ($\sim 22\text{ m}$ per pixel) showing details of a portion of the crater wall. **e**, Image ($\sim 65\text{ m}$ per pixel) highlighting the small craters with olivine annuli. **f**, Image ($\sim 22\text{ m}$ per pixel) showing details of (1) portion of the crater wall. **g**, **h**, Framing Camera images of Arruntia crater. **g**, Image ($\sim 22\text{ m}$ per pixel) showing locations of concentrations of olivine-rich material (arrows) in Arruntia ejecta. **h**, Image ($\sim 65\text{ m}$ per pixel) showing details of crater wall geology. Lenses of bright material are present (5), and dark materials are comingling with bright materials (6).



implies that mantle olivine would be excavated only by large, basin-forming impacts.

In the Bellicia–Arruntia region, we see patches of nearly pure olivine, hundreds of metres in size, in a background of howarditic material that suggest a large olivine-dominated source, with coherent sub-kilometre-size ejecta. The serial-magmatism model envisions smaller scales of petrologic variation^{14,24}, suggesting a mixed lithology of olivine and orthopyroxene that is not observed. The occurrence of several olivine spots a few hundreds of metres across, as seen in the walls of Bellicia, seems hard to reconcile with the plutonic origin.

In the magma-ocean model, the Rheasilvia basin, superimposed on the older Veneneia^{25,26} basin, could have excavated and redistributed mantle material across Vesta²⁷. The mineralogical diversity of the equatorial regions versus southern regions^{7,8} indicates that the lower crust and upper mantle, which are dominated by diogenitic material, were exposed by these impacts and were deposited as an extensive area of Rheasilvia ejecta in the northwest direction (Fig. 1), but most probably not extending to the Bellicia–Arruntia region.

Figure 3 | Spectral characteristics of the olivine-rich areas. **a**, Continuum-removed average Vesta spectrum and continuum-removed spectrum of the olivine-rich area in Bellicia. Olivine-rich spectra show a large asymmetric BI, typical of olivine-rich mixtures, whereas BII indicates that pyroxene is also present. The BI centre is at a slightly longer wavelength with respect the average spectrum, but the BII centre does not shift, as would be the case for iron-rich pyroxenes typical of eucrites or for high-calcium clinopyroxenes. **b**, Coloured lines show spectra of mixtures of olivine (Ol) and orthopyroxene (Ortho-Pyx) (data from the RELAB database) and the green points show the spectrum of the olivine-rich area. Laboratory olivine spectra exhibit only a broad, asymmetric 1-μm feature due to the overlapping of three individual absorptions²⁹, whereas orthopyroxene exhibits two well-defined, symmetric absorptions near 1 μm and, respectively, 2 μm (refs 17, 30). Spectra of mixtures of olivine–orthopyroxene show that large olivine contents (>50%) produce distortion of the band shape near 1 μm from that of pure pyroxene. More sensitive indications of olivine in a mixture are a shallow depression near 1.3 μm and a reduction in depth of BII pyroxene absorption. **c**, Scatter plot of band centres. HED meteorite data are represented as coloured circles and lie on a linear correlation trend: eucrites and diogenites data are well separated, with the howardite data between them. For olivine–orthopyroxene mixtures (30–70%, 50–50%, 70–30%; orange squares), the BI centre shifts towards longer wavelengths for increasing olivine content as illustrated by the arrow, but little or no shift is registered in the BII centre. The olivine-rich area inside Bellicia (green squares), Bellicia walls (brown cross) and a control area nearby Bellicia (cyan cross) are also represented. The olivine-rich points scatter above the HED meteorite trend and separate from the control area, which lies in the HED meteorite field. The Bellicia walls data lie between the olivine-rich area and the control area, suggesting a mixing of both.

The presence of the olivine in the hemisphere opposite the large southern basins raises the question of antipodal focusing of energy leading to excavation of olivine-rich materials from depth. However, a large, high-velocity metallic core, such as in Vesta¹, should defocus and deflect the energy away from the collision²⁸. Thus, the olivine is probably not due to antipodal excavation.

Diogenite-rich materials in the northern regions are concentrated in an area broadly corresponding to a 180-km ruined crater²⁵ near Bellicia and in other large craters farther north (Fig. 1). Thus, the northern diogenitic material might have been ejected by these other ancient large impacts. However, the depths of the old basins near Bellicia and Arruntia are 10 and 15 km (ref. 25), respectively, possibly making the basins too shallow to reach the mantle.

A generalized geologic history for these olivine-rich materials could be as follows: ancient large impacts excavated and incorporated large blocks of diogenite-rich and olivine-rich material into the eucritic crust, and subsequent impacts exposed this olivine-rich material in Arruntia and Bellicia. This produced olivine-rich terrains in a howarditic background, with diogenite-rich howardites filling nearby, eroded, older basins.

The large exposures of olivine-rich material and their association with howardite may favour a magma-ocean model for the origin of the olivine. However, the apparent absence of olivine concentrations in Rheasilvia, where the excavation depth is greater, may suggest that the internal distribution of lithologies was heterogeneous, perhaps supporting the serial-magmatism model, or that the crust–mantle boundary was deeper in the region excavated by Rheasilvia than in the Bellicia–Arruntia region. In any case, the lack of pure olivine in the southern deeply excavated basins and its unexpected discovery in the northern hemisphere of Vesta indicate a more complex evolutionary history than inferred from pre-Dawn models.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 April; accepted 13 September 2013.

Published online 6 November 2013.

1. Russell, C. T. *et al.* Dawn at Vesta: testing the protoplanetary paradigm. *Science* **336**, 684–686 (2012).

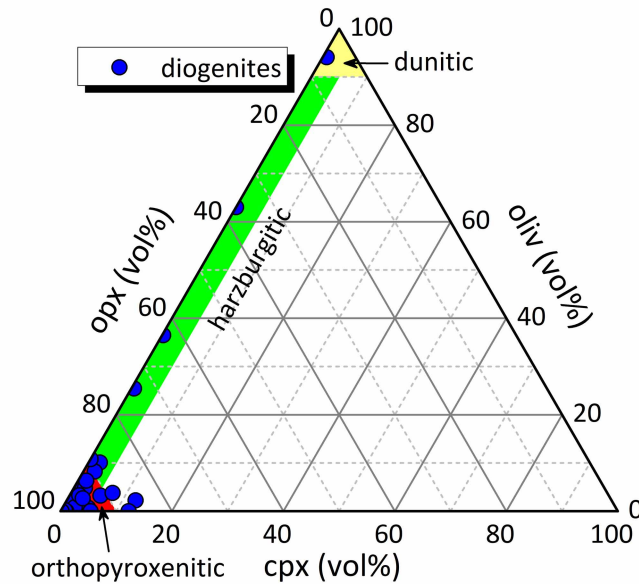
2. Beck, A. W. & McSween, H. Y. Diogenites as polymict breccias composed of orthopyroxenite and harzburgite. *Meteorit. Planet. Sci.* **45**, 850–872 (2010).
3. Gaffey, M. J. Surface lithologic heterogeneity of asteroid 4 Vesta. *Icarus* **127**, 130–157 (1997).
4. Binzel, R. P. *et al.* Geologic mapping of Vesta from 1994 Hubble Space Telescope Images. *Icarus* **128**, 95–103 (1997).
5. Li, J. Y. *et al.* Photometric mapping of asteroid (4) Vesta's southern hemisphere with Hubble Space Telescope. *Icarus* **208**, 238–251 (2010).
6. McSween, H. J. *et al.* Composition of the Rheasilvia basin, a window into Vesta's interior. *J. Geophys. Res.* **118**, 335–346 (2013).
7. De Sanctis, M. C. *et al.* Spectroscopic characterization of mineralogy and its diversity across Vesta. *Science* **336**, 697–700 (2012).
8. Ammannito, E. *et al.* Vestan lithologies mapped by the visual and infrared spectrometer on Dawn. *Meteorit. Planet. Sci.* <http://dx.doi.org/10.1111/maps.12192> (13 September 2013).
9. De Sanctis, M. C. *et al.* Vesta's mineralogical composition as revealed by the visible and infrared spectrometer on Dawn. *Meteorit. Planet. Sci.* <http://dx.doi.org/10.1111/maps.12138> (8 July 2013).
10. De Sanctis, M. C. *et al.* The VIR spectrometer. *Space Sci. Rev.* **163**, 329–369 (2011).
11. Righter, K. & Drake, M. J. A magma ocean on Vesta: core formation and petrogenesis of eucrites and diogenites. *Meteorit. Planet. Sci.* **32**, 929–944 (1997).
12. Ruzicka, A., Snyder, G. A. & Taylor, L. A. Vesta as the howardite, eucrite and diogenite parent body: implications for the size of a core and for large-scale differentiation. *Meteorit. Planet. Sci.* **32**, 825–840 (1997).
13. Barrat, J.-A. *et al.* Relative chronology of crust formation on asteroid Vesta: insights from the geochemistry of diogenites. *Geochim. Cosmochim. Acta* **74**, 6218–6231 (2010).
14. Mittlefehldt, D. W. Petrology and geochemistry of the Elephant Moraine A79002 diogenite: a genomict breccia containing a magnesian harzburgite component. *Meteorit. Planet. Sci.* **35**, 901–912 (2000).
15. Singer, R. B. Near-infrared spectral reflectance of mineral mixtures: systematic combinations of pyroxenes, olivine, and iron oxides. *J. Geophys. Res.* **86**, 7967–7982 (1981).
16. Cloutis, E. A. *et al.* Calibration of phase abundance, composition, and particle size distribution for olivine–orthopyroxene mixtures from reflectance spectra. *J. Geophys. Res.* **91**, 11641–11653 (1986).
17. Adams, J. B. Visible and near-infrared diffuse reflectance spectra of pyroxenes as applied to remote sensing of solid objects in the solar system. *J. Geophys. Res.* **79**, 4829–4836 (1974).
18. Gaffey, M. J. *et al.* Mineralogic variations within the S-type asteroid class. *Icarus* **106**, 573–602 (1993).
19. Delaney, J. S. *et al.* The polymict eucrites. *J. Geophys. Res.* **89**, C251–C288 (1984).
20. Beck, A. W. *et al.* Petrologic and textural diversity among the PCA 02 howardite group, one of the largest pieces of the Vestan surface. *Meteorit. Planet. Sci.* **47**, 947–969 (2012).
21. Beck, A. W. *et al.* MIL 03443, a dunite from asteroid 4 Vesta: evidence for its classification and cumulate origin. *Meteorit. Planet. Sci.* **46**, 1133–1151 (2011).
22. Marchi, S. *et al.* High-velocity collisions from the lunar cataclysm recorded in asteroidal meteorites. *Nature Geosci.* **6**, 303–307 (2013).
23. DeMeo, F. *et al.* An extension of the Bus asteroid taxonomy into the near-infrared. *Icarus* **202**, 160–180 (2009).
24. Shearer, C. K., Burger, P. & Papike, J. J. Petrogenetic relationships between diogenites and olivine diogenites: implications for magmatism on the HED parent body. *Geochim. Cosmochim. Acta* **74**, 4865–4880 (2010).
25. Jaumann, R. *et al.* Vesta's shape and morphology. *Science* **336**, 687–690 (2012).
26. Marchi, S. *et al.* The violent collisional history of asteroid 4 Vesta. *Science* **336**, 690–694 (2012).
27. Jutzi, M. *et al.* The structure of the asteroid 4 Vesta as revealed by models of planet-scale collisions. *Nature* **494**, 207–210 (2013).
28. Watts, A. W. *et al.* The formation of terrains antipodal to major impacts. *Icarus* **93**, 159–168 (1991).
29. Sunshine, J. M. & Pieters, C. M. Determining the composition of olivine from reflectance spectroscopy. *J. Geophys. Res.* **103**, 13,675–13,688 (1998).
30. Burns, R. G. *Mineralogical Applications of Crystal-Field Theory* (Cambridge Univ. Press, 1970).

Supplementary Information is available in the online version of the paper.

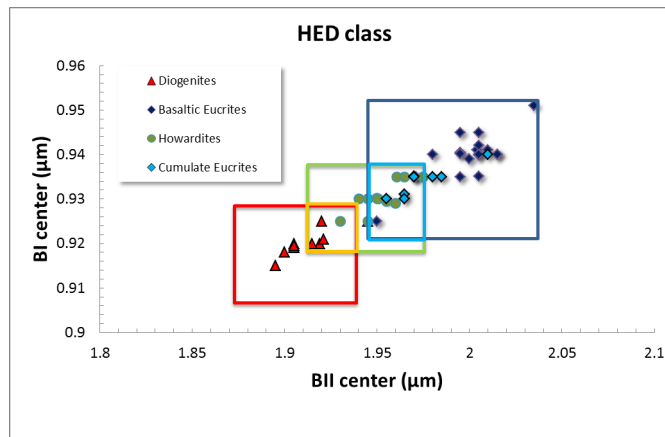
Acknowledgements We gratefully acknowledge the support of the Dawn Instrument, Operations and Science teams, and, in particular, the Dawn Framing Camera team. This work was supported by Italian Space Agency grant I/004/12/0 and by NASA through the Dawn mission and the Dawn at Vesta Participating Scientists Program.

Author Contributions M.C.D.S., E.A., E.P. and A.L. contributed to the data analysis. M.C.D.S., E.A., S.M., D.W.M., H.Y.M. and C.M.P. contributed to the data interpretation and to writing and improving the manuscript. E.A. and M.C.D.S. provided calibrated VIR data. F.T. provided geometric data. F.Z. and A.F. provided the projected and mosaicked VIR data. All authors contributed to discussion of the results.

Author Information All Dawn data are available at PDS: Small Bodies Node (http://pdssbn.astro.umd.edu/data_sb/missions/dawn/index.shtml), and VIR data are also available at the ASI Data Center (<http://www.asdc.asi.it/>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.C.D.S. (maricristina.desanctis@iaps.inaf.it) or E.A. (eleonora.ammannito@iaps.inaf.it).



Extended Data Figure 1 | Ternary diagram of orthopyroxene, olivine and clinopyroxene in diogenites. Proportions of orthopyroxene, olivine and clinopyroxene in diogenites normalized to 100%, with fields for orthopyroxenitic (red), harzburgitic (green) and dunitic diogenites (yellow). Data taken from Extended Data Table 1.



Extended Data Figure 2 | Distribution of the band centres for the HED meteorites. The difference in spectral properties of diogenites, howardites and eucrites can be quantified using a scatter plot of the BI-centre position versus the BII-centre position. We used spectra in the RELAB database to define the different HED meteorite spectral areas⁹. The HED meteorite distribution map has been derived as explained in refs 6, 8, 9. In this diagram, diogenites and eucrites populate distinct areas because both the BI-centre position and the BII-centre position are sensitive to the pyroxene compositions. Howardites, which are physical mixtures of diogenite and eucrite, plot between, and partly overlap, these fields. By associating a colour indication of composition with every region in the scatter plot (red for diogenite, green for howardite and purple for eucrite, with overlapping fields of yellow for diogenite–howardite and cyan for eucrite–howardite), we constructed the correspondence map in Fig. 1 using the same colour scheme.

Extended Data Table 1 | Average modal mineralogy of diogenites (vol%)

diogenite	ortho- pyroxene	pigeon- ite	high-Ca pyroxene	oliv- ine	plagio- clase	silica	chromite	phos- phate	troilite	metal	sum
A-881548	37.0	-	-	63.0	-	-	-	-	-	-	100.0
Aïoun el Atrouss	97.8	-	1.0	0.4	-	0.5	0.3	-	-	-	100.0
ALH 85015	85.1	-	5.2	3.0	4.8	0.4	1.0	-	0.7	-	100.2
ALHA77256	87.0	-	2.0	10.0	-	-	0.9	-	0.1	-	100.0
EET 83246	94.1	-	4.6	0.7	-	0.5	0.2	-	0.1	-	100.2
EET 83247	96.8	-	0.6	0.2	-	1.8	0.3	-	0.1	0.2	100.0
EET 87530	93.1	-	1.2	-	-	0.6	5.1	-	0.1	0.1	100.2
EETA79002	61.8	-	0.2	35.6	0.1	0.1	0.2	-	1.9	0.1	100.0
Ellemeet	86.9	-	1.4	4.7	-	0.1	2.9	-	3.3	0.8	100.1
Garland	92.6	-	1.0	0.1	0.4	0.1	4.8	-	0.6	0.4	100.0
GRA 98108	73.7	-	0.5	25.3	0.2	0.1	0.1	-	0.2	0.1	100.2
Ibbenbüren	92.6	-	5.3	-	1.0	0.5	0.7	-	-	-	100.1
Johnstown	95.7	-	0.5	-	0.6	0.5	0.1	-	2.6	0.2	100.2
LAP 02216	96.7	-	0.3	2.0	-	-	0.2	-	0.6	0.2	100.0
LAP 03979	89.0	-	1.9	4.6	3.9	-	0.2	-	0.3	0.1	100.0
LAP 91900	98.2	-	0.2	-	-	0.9	0.1	-	0.3	0.4	100.1
LEW 88008	94.1	-	2.1	0.8	2.1	0.4	0.3	-	0.3	0.1	100.2
LEW 88679	88.4	-	2.0	8.1	0.6	-	0.1	-	0.8	0.1	100.1
Manegaon	95.0	-	1.0	-	3.7	-	0.1	-	0.1	-	99.9
MET 01084	88.6	-	1.4	6.1	3.6	-	0.1	-	0.1	-	99.9
MIL 03368	83.5	-	12.1	2.2	-	0.2	1.8	-	0.1	0.1	100.0
MIL 03443	5.0	-	0.7	91.0	-	-	1.0	-	2.0	0.3	100.0
MIL 07001	89.2	-	-	10.8	-	-	-	-	-	-	100.0
PCA 02008	86.8	-	7.3	3.7	1.3	0.3	0.3	-	0.2	0.1	100.0
PCA 91077	97.6	-	0.9	-	-	1.1	0.4	-	-	-	100.0
Peckelsheim	92.7	-	1.7	3.2	0.7	0.3	1.4	-	0.1	-	100.1
QUE 99050	97.8	-	0.7	-	-	-	0.6	0.9	-	-	100.0
Roda	92.5	-	2.6	2.6	1.8	0.2	0.1	-	0.2	-	100.0
Shalka	98.0	-	0.3	-	-	0.2	1.4	-	0.3	-	100.2
Tatahouine	99.6	-	-	-	-	0.3	0.1	-	-	-	100.0
TIL 82410	97.0	-	0.7	0.2	-	-	2.1	-	-	-	100.0
Y-74011	98.4	-	0.2	-	-	0.5	-	-	0.9	0.1	100.1
Yamato Type B	62.1	16.9	6.2	-	12.1	1.5	0.5	-	0.4	-	99.7

Average modal mineralogy of diogenites compiled from different literature sources (refs 2, 19, 21 and refs 33–35 in Supplementary Information).

Late-twentieth-century emergence of the El Niño propagation asymmetry and future projections

Agus Santoso¹, Shayne McGregor¹, Fei-Fei Jin², Wenju Cai³, Matthew H. England¹, Soon-Il An⁴, Michael J. McPhaden⁵ & Eric Guilyardi^{6,7}

The El Niño/Southern Oscillation (ENSO) is the Earth's most prominent source of interannual climate variability, exerting profound worldwide effects^{1–7}. Despite decades of research, its behaviour continues to challenge scientists. In the eastern equatorial Pacific Ocean, the anomalously cool sea surface temperatures (SSTs) found during La Niña events and the warm waters of modest El Niño events both propagate westwards, as in the seasonal cycle⁷. In contrast, SST anomalies propagate eastwards during extreme El Niño events, prominently in the post-1976 period^{7–10}, spurring unusual weather events worldwide with costly consequences^{3–6,11}. The cause of this propagation asymmetry is currently unknown¹⁰. Here we trace the cause of the asymmetry to the variations in upper ocean currents in the equatorial Pacific, whereby the westward-flowing currents are enhanced during La Niña events but reversed during extreme El Niño events. Our results highlight that propagation asymmetry is favoured when the westward mean equatorial currents weaken, as is projected to be the case under global warming^{12–14}. By analysing past and future climate simulations of an ensemble of models with more realistic propagation, we find a doubling in the occurrences of El Niño events that feature prominent eastward propagation characteristics in a warmer world. Our analysis thus suggests that more frequent emergence of propagation asymmetry will be an indication of the Earth's warming climate.

The tropical Pacific is home to intense convection, allowing for strong thermal and dynamical interactions between the upper ocean and the overlying atmosphere¹⁵. As warm SST anomalies propagate eastwards during extreme El Niño events (for example, 1982–83, 1997–98; see Extended Data Fig. 1a), nonlinear dynamical heating processes tend to intensify the anomalously warm SST⁹, while the western Pacific warm pool (water with temperature exceeding 28 °C) extends eastwards, moving the eastern edge of the warm pool beyond 160° W. This induces an eastward shift of equatorial rainfall and an extreme swing of the Southern Hemisphere's largest rainband, the South Pacific Convergence Zone¹¹, causing extreme hydroclimatic conditions that most severely affect vulnerable island countries in the Pacific^{11,16}. Beyond the Pacific, almost every continent felt the impacts of the drastic shift in weather patterns during the 1982–83 extreme El Niño event, and in the USA alone crop losses were estimated to be around \$10–12 billion⁴ (approximately \$24–26 billion in 2013 dollars).

These profound impacts demand an improved understanding of ENSO propagation dynamics. Many studies have evaluated the relative importance of various ocean–atmosphere feedback processes^{17–19}, yet the mechanism for the propagation asymmetry remains unresolved. Here we show that an asymmetry in the zonal flow along the equatorial Pacific upper ocean (hereafter referred to as the equatorial Pacific current) is the main cause.

Using various observational data assimilation systems (see Methods and Supplementary Table 1), we quantify the propagation characteristics of temperature anomalies (T^a) by compositing the equatorial warming and cooling rates (the time derivative of temperature, dT/dt) of the surface mixed layer for the strongest El Niño events on record (1982–83 and 1997–98; see Fig. 1a) and all La Niña events (Fig. 1b) since 1976. The composite covers the evolution over a two-year period, before and after the event peak (usually in January, denoted 'Jan (1)'; see Fig. 1 legend). The contour of $dT/dt=0$ marks the peak of the temperature anomaly (dashed curve in Fig. 1a and b). A linear regression using the samples of zero-value rates is constructed (green line). The slope β describes the propagation characteristics: a positive slope indicating temperature anomalies peak earlier in the west, that is, eastward propagation; a negative slope indicating westward propagation; and the greater the amplitude, the slower the propagation. This analysis shows opposite zonal propagation of SST anomalies between these two types of events; eastwards during extreme El Niño events ($\beta = 0.82$) and westwards during La Niña events ($\beta = -0.46$). During moderate El Niño events the propagation is westwards (Extended Data Fig. 1b), similar to La Niña events.

The direction of propagation has been understood as arising from three main competing positive feedback processes^{17,18}. The zonal advective and Ekman pumping feedbacks associated with fluctuations in the trade winds involve advection of climatological SST along the Equator by anomalous zonal currents (u^a) and upwelling (w^a)^{20,21}, respectively (that is, $u^a(dT/dx)$ and $w^a(dT/dz)$; where the overbar indicates climatological mean and superscript 'a' indicates anomaly. The thermocline feedback involves vertical advection ($\bar{w}(dT/dz)^a$) associated with eastward-propagating internal waves that influence SST in the eastern Pacific through the mean upwelling (\bar{w}) of water at the thermocline (a narrow depth range of strong vertical temperature gradients below the mixed layer). These processes can establish propagation of SST anomalies in either direction: eastward if the thermocline feedback dominates, and westward otherwise¹⁷. Linear theories have highlighted a higher importance of the thermocline feedback in the decades since the mid-1970s^{8,22,23}, however, this would predict an eastward propagation during La Niña events as well¹⁹, in contrast to observations¹⁰ (Fig. 1b).

La Niña anomalies can be viewed as an enhancement of the prevailing climate, with stronger westward-flowing surface currents. On the other hand, eastward current anomalies during El Niño associated with anomalously weak trade winds²⁴ (Extended Data Fig. 2), oppose and even exceed in amplitude the background current, leading to a net eastward flow (Fig. 1c). This asymmetry in the total zonal current is apparent in all reanalyses (Extended Data Fig. 3, Supplementary Tables 2 and 3). Our heat budget analysis (see Methods) shows that advection of anomalous temperature by the total current, $(\bar{u} + u^a)(dT/dx)^a$,

¹Australian Research Council (ARC) Centre of Excellence for Climate System Science and Climate Change Research Centre, Level 4 Mathews Building, The University of New South Wales, Sydney 2052, Australia. ²Department of Meteorology, School of Ocean and Earth Science and Technology (SOEST), University of Hawaii at Manoa, 2525 Correa Road, Honolulu, Hawaii 96822, USA. ³Commonwealth Scientific and Industrial Research Organisation (CSIRO), Marine and Atmospheric Research, 107–121 Station Street, Aspendale, Victoria 3195, Australia. ⁴Department of Atmospheric Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-Gu, Seoul 120-749, South Korea. ⁵National Oceanic and Atmospheric Administration (NOAA)/Pacific Marine Environmental Laboratory, Seattle, Washington 98115, USA. ⁶Laboratoire d'Océanographie et du Climat: Expérimentation et Approches Numériques/Institut Pierre Simon Laplace (IPSL)/Centre national de la recherche scientifique (CNRS), tour 45-55, étage 4, pièce 406, Université Pierre et Marie Curie, 4 Jussieu, 75252 Paris Cedex 05, France. ⁷National Centre for Atmospheric Science (NCAS)—Climate, Department of Meteorology, University of Reading, Earley Gate, Reading RG6 6BB, UK.

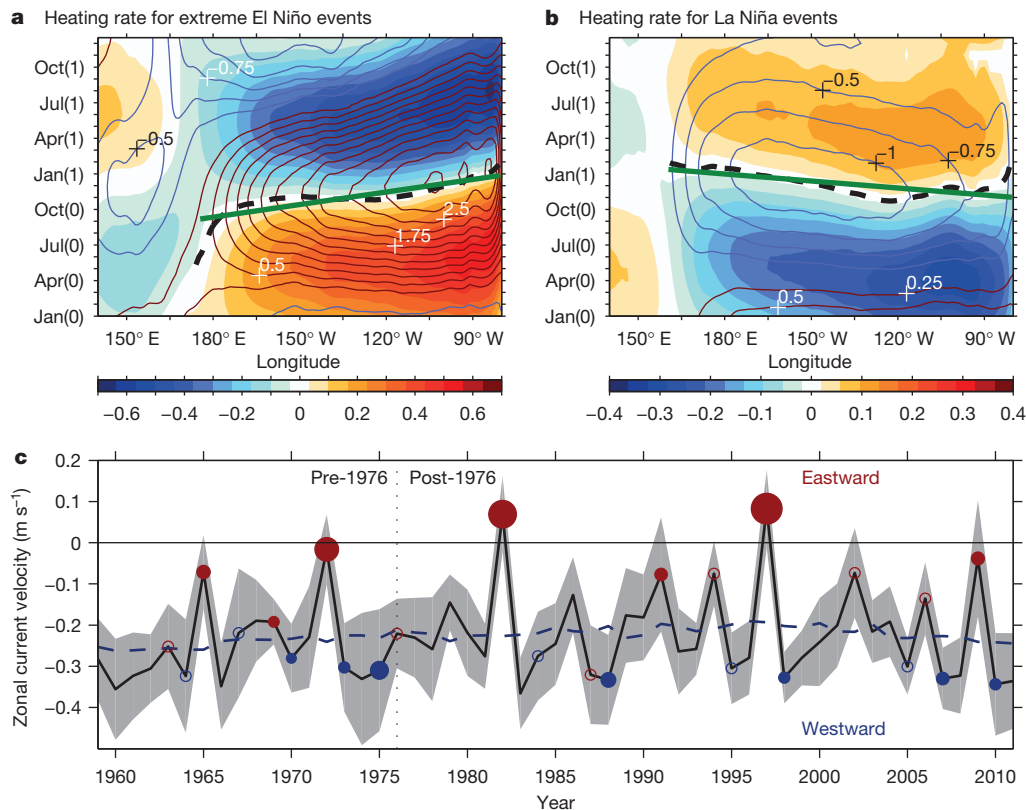


Figure 1 | Equatorial Pacific current and zonal propagation of ENSO SST anomalies. **a**, Warming and cooling rate (colour scale; in units of °C per month) of the equatorial Pacific surface mixed layer on interannual timescales (an average over 5° S–5° N) composited over an extreme El Niño (1982, 1997) lifecycle across all reanalysis products. Red (or blue) contours indicate the associated positive (or negative) SST anomalies. Statistically significant values at the 95% confidence level are shaded and contoured. The monthly composites are shown over the developing and decaying ENSO phases, denoted as year 0 and year 1, respectively, in which the events typically peak in January of year 1 ('Jan(1)'). **b**, As in **a** but for post-1976 La Niña events. The mean of the linear-fit slope (green line) of the phase transition (dashed black line), β (in units of

seconds per metre), across the products is 0.82 in **a** and -0.45 in **b**, both significant above the 95% confidence level ($P < 0.05$) based on a bootstrap method with 1,000 draws. **c**, Zonal current velocity averaged across the reanalysis products, over 5° S–5° N, 160° E–90° W, and over August to December. The dashed curve highlights interdecadal variation using a 13-year running mean. Grey shading denotes two standard deviations about each mean value, representing monthly spread and variations across reanalyses. Red and blue filled circles indicate occurrences of strong El Niño and La Niña events, respectively, with relative event intensity indicated by different marker sizes. Open circles indicate moderate ENSO events.

which is strongly westward during La Niña events but eastward during extreme El Niño events (Fig. 1c; Extended Data Fig. 4), represents one salient asymmetric feature that should be considered.

We therefore examine how the mixed-layer heat balance changes when the total current-induced heat flux $(\bar{u} + u^a)(dT/dx)^a$ is removed from the heat-budget equation (see Methods). Evolution of the residual warming and cooling rates shows that without the effect of the total current, an eastward propagation would result for extreme El Niño and La Niña events (red dashed line in Figs 2a and b), both with a positive slope of β^* , 0.55 and 0.69, respectively. This reconstructs a linear framework in which the thermocline feedback dominates, leading to an eastward propagation for all events after 1976 (ref. 19). Thus it is the westward total flow that plays a key part in determining the westward propagation during La Niña events.

The role of the total current can be further understood by decomposing it into one component that is due to the long-term mean current $\bar{u}(dT/dx)^a$ and another due to the ENSO-related current anomaly $u^a(dT/dx)^a$. We find that the westward long-term mean current favours a westward propagation during both types of events; without it the eastward propagation during extreme El Niño events is more prominent and the propagation during La Niña events reverses to eastward (red dashed line, Extended Data Fig. 1d and e). The current anomaly, on the other hand, has an opposite effect on the two types of events (Supplementary Table 4). Without the effect of the eastward current anomaly, the eastward tendency during extreme El Niño events is severely weakened

(red dashed line in Fig. 2c). This eastward anomalous current, stronger than the westward mean current, leads to a flow reversal during extreme El Niño events (Fig. 1c), making the eastward propagation characteristic more prominent (Fig. 2a). During La Niña events, on the other hand, the anomalous current reinforces the effect of the westward mean flow for a more pronounced westward propagation (Fig. 2b and d and Extended Data Fig. 1e). During moderate El Niño events, the eastward current anomaly is far weaker than the mean current, and thus the total current remains westward (Fig. 1c, Extended Data Fig. 1c). Thus, for moderate ENSO events, there is no asymmetry, and SST anomalies for both El Niño and La Niña events propagate westwards (Supplementary Table 4).

The nonlinear effect is more prominent post-1976 (Supplementary Table 4) when El Niño events are stronger^{9,25} with large eastward current anomalies that are occasionally comparable to, or greater than, the mean background current (Fig. 1c). During such events, this effect reinforces eastward propagation induced by the thermocline feedback. During a La Niña event, the westward current, along with the zonal advective and Ekman pumping feedbacks, weakens the thermocline feedback effect, resulting in a net westward propagation (Fig. 3). This superposition of ENSO-related large current anomalies onto the long-term mean westward current invalidates the assumptions of linearity, making linear theories unable to explain the propagation asymmetry.

Thus, interplay between the ENSO-related current anomaly and the climatological current determines the way the equatorial Pacific circulation influences the zonal propagation of SST anomalies. This means

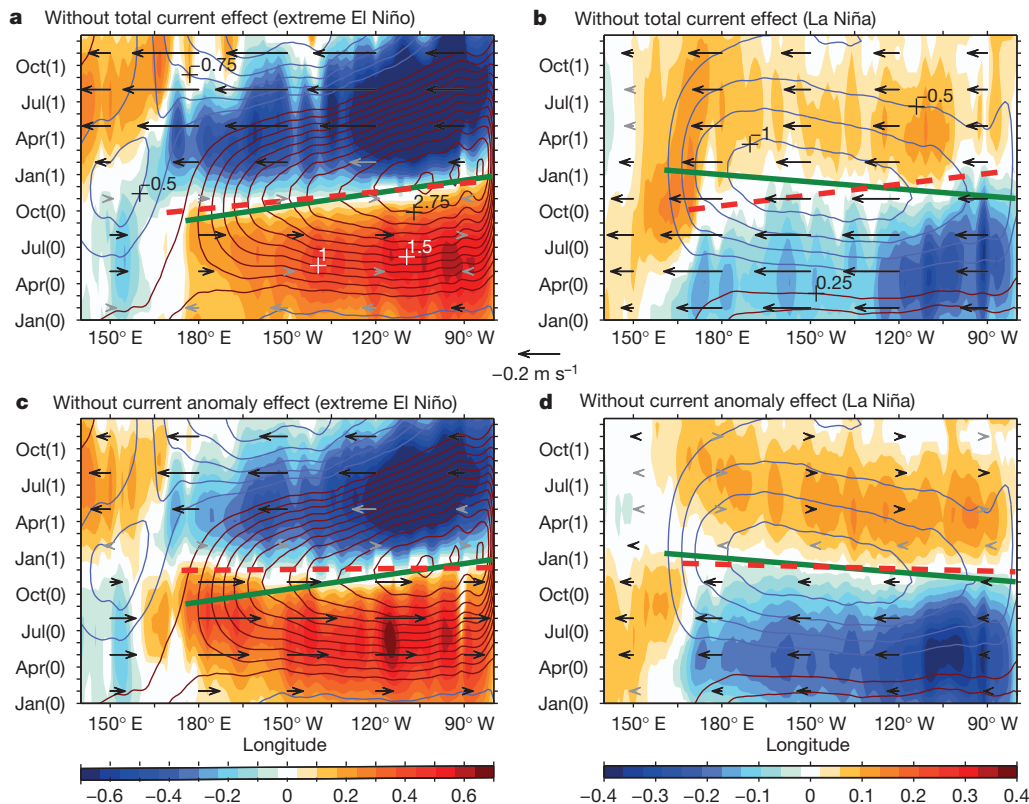


Figure 2 | Effect of total and anomalous currents on equatorial Pacific mixed layer heat balance during extreme El Niño events and all La Niña events. **a**, Composite evolution of the interannual-scale mixed layer warming and cooling rates (colour scale; in units of °C per month) during extreme El Niño events (1982, 1997), with advection of temperature anomalies by the total current (arrow) removed. The red dashed line and the green line indicate the altered slope β^* and the original β , respectively. **b**, As in **a** but for post-1976

La Niña events. Only statistically significant values at the 95% confidence level are shaded in colour, contoured or marked by black arrows (grey arrows otherwise). β^* in **a** and **b** are respectively 0.55 and 0.69, both significant above the 95% confidence level ($P < 0.05$). **c**, As in **a** but with the effect of current anomaly removed. **d**, As in **c** but for La Niña events. β^* in **c** and **d** are respectively 0.05 and -0.14 , both not statistically significant ($P > 0.4$).

that a change in ENSO intensity or in the mean current can influence the extent to which the propagation asymmetry can be observed. The post-1976 prominence of the propagation asymmetry is partly because of the extremity of the 1982–83 and 1997–98 El Niño events. The mean current itself weakened through the 1980–2000 period (dashed curve in Fig. 1c), with a consistent weakening of the trade winds (Extended Data Fig. 2b). Although this mean current reduction could be interpreted as a rectification of a change in ENSO variability^{26,27}, a weakened mean current will favour occurrences of an eastward propagation, even if El Niño intensity does not change. At present, there is no agreement among climate models on the magnitude of future ENSO events^{28,29}. However, the consensus that emerges is a future with weaker equatorial mean westward currents^{12–14}. Our study implies that this would increase the likelihood for occurrences of El Niño events with prominent eastward propagation characteristics.

To this end, we analysed 40 climate models that participated in the Coupled Model Intercomparison Project phases 3 and 5 (CMIP3 and CMIP5), subject to increasing atmospheric CO₂ concentration (see Methods). With the large number of simulated ENSO events, the multi-model aggregate demonstrates robust statistics reaffirming the above conclusion that weaker mean currents and current reversals, which are projected to be more typical in the future, facilitate eastward propagation (see Methods and Extended Data Figs 5–7). Indeed, we find that a subset of models that are more realistic in terms of flow features and frequency of El Niño events with prominent eastward propagation (19 models; see Methods and Extended Data Figs 8–10 for model selection) simulate a 100% aggregate increase in the mean occurrence of such El Niño events (Fig. 4a), from about 2.7 events in 1907–1999 to about 5.4 events in 2006–2098.

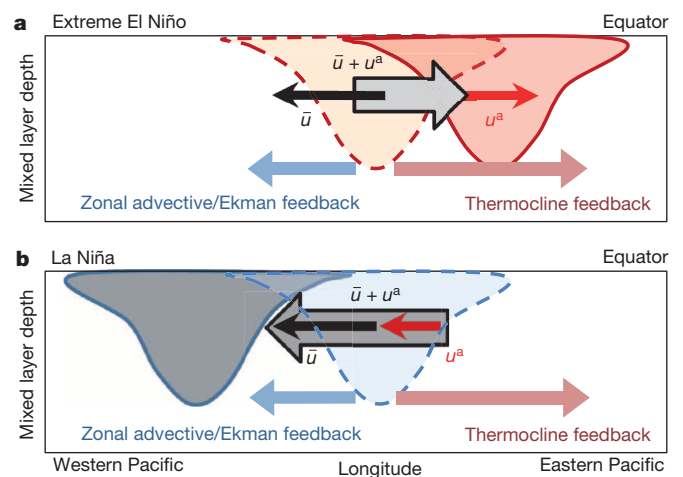


Figure 3 | Schematic of competing effects on zonal propagation direction during ENSO events. **a**, Zonal currents in the equatorial Pacific (large grey arrow) have the effect of shifting the initial warm surface anomalies (dashed red patch) eastwards during extreme El Niño events, because the current anomaly u^a (red arrow) is eastward and exceeds the strength of the westward background current \bar{u} (black arrow). This effect counters westward propagation as induced by the zonal advective and Ekman pumping feedbacks (blue arrow) and enhances eastward propagation induced by the thermocline feedback (pink arrow). **b**, During La Niña events, the zonal currents are prominently westward because the current anomaly always enhances the westward-flowing mean current. This weakens the thermocline feedback effect and enhances westward propagation as induced by the other two dynamical feedbacks.

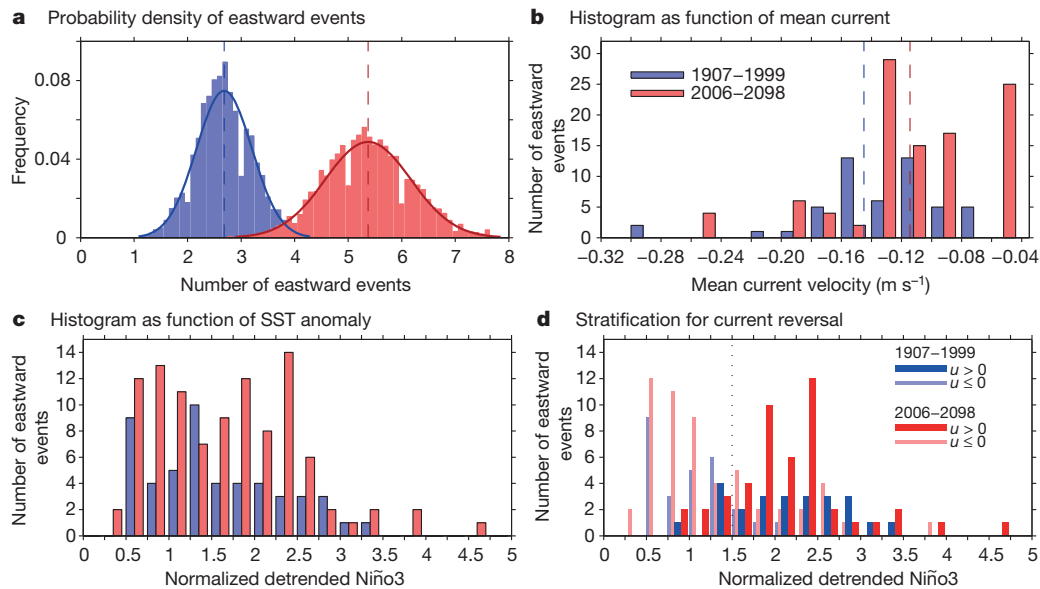


Figure 4 | Statistics of El Niño occurrence characterized by prominent eastward propagation in CMIP3 and CMIP5 models. **a**, Probability density function of 19 model ensemble means in the past (1907–1999; blue) and future (2006–2098; red) periods. Each probability distribution is generated from 5,000 bootstrap draws using the 19 model samples. The solid curves overlaid are the corresponding probability density functions of a fitted normal distribution, which are significantly different above the 95% confidence level ($P = 0.01$). Dashed red and blue vertical lines indicate the corresponding

The role of the current is further highlighted as the increase in eastward-propagating events is skewed towards weaker mean currents (Fig. 4b; Extended Data Fig. 10b), and this occurs for El Niño events of all magnitude (Fig. 4c). There is, in particular, a tendency for the increase to be larger in models that project stronger ENSO amplitude (Extended Data Fig. 10c), which is in turn associated with more occurrences of a current reversal (Extended Data Fig. 10d), a feature unique to the 1982–83 and 1997–98 El Niño events.

Stratifying the statistics (Fig. 4c) in terms of a current reversal or otherwise, 45% of the increase is found to be associated with current reversals, most of which are El Niño events stronger than the typical magnitude of past events (Fig. 4d). However, the inter-model consensus for ENSO amplitude projection is weak, despite a reduced mean current in all models (Extended Data Fig. 10b and c). This suggests that a weakened mean current is the determining factor for future increases in eastward propagating events of all magnitudes, including extreme El Niño events, either through the thermocline feedback effect or a current reversal, or both.

In summary, although different factors have been proposed to explain various nonlinear characteristics of ENSO^{9,25,30}, none have been found to explain the cause for its propagation asymmetry, as observed in recent decades. Here we have provided observational and modelling evidence that the equatorial Pacific current is an important element of this asymmetry. The superposition of a current anomaly during ENSO onto the long-term mean westward flow enhances the westward currents during a La Niña event, but reverses the currents during extreme El Niño events. The role of the equatorial currents highlighted here casts a fresh perspective on the fundamentals of ENSO behaviour. Given the projected weakening of the background mean flow under global warming, our analysis not only resolves a perplexing scientific issue, but suggests that increased occurrences of ENSO propagation asymmetry will be a manifestation of global greenhouse warming, with important socio-economic consequences.

METHODS SUMMARY

The propagation tendency of temperature anomalies during ENSO events is quantified as the slope of the zero-value contour of warming and cooling rates on ENSO

ensemble average. **b**, Multi-model histogram for the event occurrence as a function of long-term averaged equatorial Pacific current velocity in the past (blue) and future (red) periods, segregated into bins of 0.04 m s^{-1} . **c**, As in **b** but for the Niño3 index, which is detrended and normalized by the standard deviation of the past period, segregated into bins of size 0.25. **d**, As in **c** but stratified according to the concurrence with (thick dark bars) and without (thin light bars) current reversals. The dotted vertical line in **d** marks 1.5 units of the normalized value.

timescales that tracks the peak of temperature anomalies as they evolve in time along the equatorial Pacific (Fig. 1a, b). A positive (or negative) slope in the time-longitude space indicates eastward- (or westward-) propagating temperature anomalies: the steeper the slope the slower the zonal propagation and thus the more observable the propagation characteristic. To investigate the factors that can cause temperature anomalies to propagate zonally, we conduct a heat budget analysis of the ocean surface mixed layer. All variables are derived from five ocean reanalysis systems that assimilate high-quality observational products going back to 1980 or earlier (Supplementary Table 1). Our surface heat balance explicitly expresses the zonal advection of temperature anomaly by the mean current and is considered to interact with the nonlinear component (that is, advection by anomalous current). Removing certain heat-flux components from the heat balance would alter the slope of the zero-value contour, and so by comparing the altered slope (β^*) with the original (β) its influence on the propagation can be inferred. Because this study concerns asymmetry between El Niño and La Niña, a composite approach is adopted (see Methods for classification of ENSO events). The implication of our results for future climate is assessed through the analysis of 40 CMIP3 and CMIP5 climate models (see full Methods).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 January; accepted 20 September 2013.

Published online 17 November 2013.

- McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. ENSO as an integrating concept in Earth science. *Science* **314**, 1740–1745 (2006).
- Lehodey, P., Bertignac, M., Hampton, J., Lewis, A. & Picaut, J. El Niño/Southern Oscillation and tuna in the western Pacific. *Nature* **389**, 715–718 (1997).
- Bove, M. C., O'Brien, J. J., Eisner, J. B., Landsea, C. W. & Niu, X. Effect of El Niño on U.S. landfalling hurricanes, revisited. *Bull. Am. Meteorol. Soc.* **79**, 2477–2482 (1998).
- Wilhite, D. A., Wood, D. A. & Meyer, S. J. in *Climate Crisis* (eds Glantz, M., Katz, R. & Krenz, M.) 75–78 (UNEP, 1987).
- Changnon, S. A. Impacts of 1997–98 El Niño generated weather in the United States. *Bull. Am. Meteorol. Soc.* **80**, 1819–1827 (1999).
- Liu, Z. & Alexander, M. Atmospheric bridge, oceanic tunnel, and global climatic teleconnections. *Rev. Geophys.* **45**, RG2005, <http://dx.doi.org/10.1029/2005RG000172> (2007).
- Wallace, J. M. *et al.* On the structure and evolution of ENSO-related climate variability in the tropical Pacific: lessons from TOGA. *J. Geophys. Res.* **103**, 14241–14259 (1998).

8. Wang, B. & An, S.-I. A mechanism for decadal changes of ENSO behaviour: roles of background wind changes. *Clim. Dyn.* **18**, 475–486 (2002).
9. An, S.-I. & Jin, F.-F. Nonlinearity and asymmetry of ENSO. *J. Clim.* **17**, 2399–2412 (2004).
10. McPhaden, M. J. & Zhang, X. Asymmetry in zonal phase propagation of ENSO sea surface temperature anomalies. *Geophys. Res. Lett.* **36**, L13703, <http://dx.doi.org/10.1029/2009GL038774> (2009).
11. Cai, W. *et al.* More extreme swings of the South Pacific convergence zone due to greenhouse warming. *Nature* **488**, 365–369 (2012).
12. Vecchi, G. A. *et al.* Weakening of tropical Pacific atmospheric circulation due to anthropogenic forcing. *Nature* **441**, 73–76 (2006).
13. DiNezio, P. *et al.* Climate response of the equatorial Pacific to global warming. *J. Clim.* **22**, 4873–4892 (2009).
14. Sen Gupta, A., Ganachaud, A., McGregor, S., Brown, J. N. & Muir, L. Drivers of the projected changes to the Pacific Ocean equatorial circulation. *Geophys. Res. Lett.* **39**, L09605, <http://dx.doi.org/10.1029/2012GL051447> (2012).
15. Graham, N. E. & Barnett, T. P. Sea surface temperature, surface wind divergence, and convection over tropical oceans. *Science* **238**, 657–659 (1987).
16. Vincent, E. M. *et al.* Interannual variability of the South Pacific Convergence Zone and implications for tropical cyclone genesis. *Clim. Dyn.* **36**, 1881–1896 (2011).
17. Neelin, D. J. *et al.* ENSO theory. *J. Geophys. Res.* **103**, 14261–14290 (1998).
18. Jin, F.-F. & Neelin, J. D. Modes of interannual tropical ocean-atmosphere interaction—a unified view. Part I: Numerical results. *J. Atmos. Sci.* **50**, 3477–3503 (1993).
19. Fedorov, A. & Philander, S. G. H. A stability analysis of tropical ocean-atmosphere interactions: bridging measurements and theory for El Niño. *J. Clim.* **14**, 3086–3101 (2001).
20. An, S.-I., Jin, F.-F. & Kang, I.-S. The role of zonal advection feedback in phase transition and growth of ENSO in the Cane-Zebiak model. *J. Met. Soc. Jpn* **77**, 1151–1160 (1999).
21. Kang, I.-S., An, S.-I. & Jin, F.-F. A systematic approximation of the SST anomaly equation for ENSO. *J. Met. Soc. Jpn* **79**, 1–10 (2001).
22. An, S.-I. & Jin, F.-F. An eigenanalysis of the interdecadal changes in the structure and frequency of ENSO mode. *Geophys. Res. Lett.* **27**, 2573–2576 (2000).
23. Fedorov, A. & Philander, S. G. H. Is El Niño changing? *Science* **288**, 1997–2002 (2000).
24. Siedel, H. & Giese, B. S. Equatorial currents in the Pacific Ocean 1992–1997. *J. Geophys. Res.* **104**, 7849–7863 (1999).
25. Jin, F.-F., An, S.-I., Timmermann, A. & Zhao, J. Strong El Niño events and nonlinear dynamical heating. *Geophys. Res. Lett.* **30**, 1120, <http://dx.doi.org/10.1029/2002GL016356> (2003).
26. Choi, J., An, S.-I., Dewitte, B. & Hsieh, W. W. Interactive feedback between the tropical Pacific decadal oscillation and ENSO in a coupled general circulation model. *J. Clim.* **22**, 6597–6611 (2009).
27. Liang, J., Yang, X.-Q. & Sun, D.-Z. The effect of ENSO events on the tropical Pacific mean climate: insights from an analytical model. *J. Clim.* **25**, 7590–7606 (2012).
28. Guilyardi, E. El Niño-mean state-seasonal cycle interactions in a multi-model ensemble. *Clim. Dyn.* **26**, 329–348 (2006).
29. Collins, M. *et al.* The impact of global warming on the tropical Pacific Ocean and El Niño. *Nature Geosci.* **3**, 391–397 (2010).
30. Frauen, C. & Dommenget, D. El Niño and La Niña amplitude asymmetry caused by atmospheric feedbacks. *Geophys. Res. Lett.* **37**, L18801 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling and all modelling groups for making the CMIP data available. We thank F. Avila and J. Kajtar for downloading and processing the climate models data sets. A.S., S.M. and M.H.E. are supported by the Australian Research Council. W.C. is supported by the Australian Climate Change Science Programme. S.-I.A. was supported by the National Research Foundation of Korea funded by the Korean government (MEST) (grant number NRF-2009-C1AAA001-2009-0093042). M.J.M. is supported by NOAA. This is PMEL contribution number 3977.

Author Contributions A.S. and S.M. conceived the study in discussion with F.-F.J. A.S. designed and conducted the analysis. W.C. and A.S. wrote the initial draft of the paper. All authors contributed to interpreting results, presentation, and improvement to the paper.

Author Information The NOAA SST and GODAS reanalysis data are provided by the NOAA/OAR/ESRL PSD via <http://www.esrl.noaa.gov/psd/>. All other reanalysis data were downloaded from the Asia-Pacific Data-Research Center of the IPRC at <http://apdrc.soest.hawaii.edu/data/>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (a.santoso@unsw.edu.au).

METHODS

Heat budget analysis. We consider the heat balance of the surface mixed layer, which can be expressed as follows:

$$\begin{aligned} dT/dt = - \left[(\bar{u} + u^a)(dT/dx)^a + u^a \overline{dT/dx} + \bar{w}(dT/dz)^a + \right. \\ \left. w^a \overline{dT/dz} + d(\bar{v}T^a)/dy + v^a \overline{dT/dy} \right] + \text{Res.} \end{aligned} \quad (1)$$

The variables T , u , v and w are potential temperature, and the zonal, meridional and vertical ocean current velocities respectively. Differential operators, x , y , z , and t , are along zonal, meridional, and vertical directions, and time, respectively. Superscript 'a' and overbar denote anomalous and long-term averaged quantities, respectively. All variables are averaged between 5°S – 5°N , over the surface layer depth of 50 m. The rate of change of the mixed layer temperature (dT/dt) is calculated as monthly increments using a centred-difference approximation. Terms not explicitly expressed in equation (1), such as eddy effects and damping by net air–sea heat fluxes, are absorbed into 'Res.', such that the left- and right-hand sides of equation (1) are identical. Equation (1) is slightly different to that adopted in previous studies^{8,20,21}, in that $\bar{u}(dT/dx)^a$ is expressed explicitly here and is considered to interact with the nonlinear advection term $u^a(dT/dx)^a$. This combination $u(dT/dx)^a$, where $u = \bar{u} + u^a$, is simply interpreted as the zonal advection of temperature anomalies by the total equatorial Pacific zonal current, which can be readily observed. The term $\bar{u}(dT/dx)^a$ tends to be overlooked because it is convolved into the continuity component via volume conservation—that is, $T^a \left(\frac{d\bar{u}}{dx} + \frac{d\bar{v}}{dy} + \frac{d\bar{w}}{dz} \right) = 0$ —when the heat budget is expressed in flux form: $d(\bar{u}T^a)/dx + d(\bar{v}T^a)/dy + d(\bar{w}T^a)/dz$.

Quantification of propagation characteristic. The contour of $dT/dt = 0$ marks the peak of T^a , so its positive (or negative) slope in time–longitude space indicates eastward- (or westward-) propagating T^a (Fig. 1a, b). The phase transition slope, β , is calculated by fitting a line via the least-squares method to the contour between 160°E – 80°W and May(0)–May(1) to allow some room for temporal movement upon removal of the advection terms. The rationale for the longitudinal extent is described below under the heading 'Mean currents and ENSO propagation structures across models', and our results are not sensitive to this aspect of the calculation.

Removing an important advection component from the right-hand side of equation (1) would alter the spatial and temporal structure of dT/dt , thus affecting β . For example, a reversal from a negative slope (that is, westward propagation) to a positive slope (that is, eastward propagation) would suggest that the component removed is crucial in setting the westward propagation. In this way, the role of a certain advection term on the propagation tendency of T^a can be determined by comparing the altered slope β^* to the original β . The 95% regression standard error for the slopes is considered in all analysis by setting any slopes to zero if they are not greater than their corresponding error estimates.

Data sets and data processing. The reanalysis products used are ECMWF ORA-S3³¹, ECMWF ORA-S4³², SODA-2.16³³, SODA-2.2.4³⁴ and GODAS³⁵ (Supplementary Table 1). Each reanalysis system assimilates the available observations (such as hydrographic profile data, moorings and satellites) into an ocean model forced by observed surface wind stress to calculate ocean currents. The reanalysis systems use different ocean models and data assimilation techniques. To focus on processes at ENSO timescales, a Butterworth low-pass filter³⁶ is applied before analysis to remove signals with periods shorter than 18 months. Without filtering, the spatio-temporal structure of the warming and cooling rate dT/dt is noisy, given large high-frequency monthly fluctuations. On ENSO timescales, the rate of warming and cooling smoothly tracks the evolution of SST anomalies.

ENSO classification and statistical significance test. The classification of ENSO events is based on the Niño3 index derived from the National Oceanic and Atmospheric Administration (NOAA) extended reconstructed SST version-3b³⁷, averaged over December–February, when ENSO events typically peak. ENSO events are defined if the Niño3 amplitude, within each of the pre-1976 (1959–1976) and post-1976 (1976–2011) periods, is greater than 0.5 units of standard deviation. We classify these as strong if Niño3 exceeds 1 unit of standard deviation, and as moderate or weak otherwise. This yields the following classification of events (the developing phase year is quoted): strong El Niño events in 1965, 1969, 1972, 1982, 1991, 1997 and 2009; strong La Niña events in 1970, 1973, 1975, 1988, 1998, 2007 and 2010; moderate El Niño events in 1963, 1976, 1987, 1994, 2002 and 2006; moderate La Niña events in 1964, 1967, 1984, 1995 and 2005. The 95% statistical significance for each composite is evaluated using a bootstrap approach³⁸ in which samples of size N are randomly drawn repeatedly to obtain 1,000 mean values. N is the number of ENSO events within each respective period pooled together for all the reanalysis products. All significance levels are evaluated based on the two-sided P -value.

Analysis of climate models. The observational analysis results demonstrate that, in the backdrop of the effects by the three ENSO dynamical feedbacks, the

equatorial Pacific current is an important element for the zonal phase propagation of ENSO SST anomalies (Fig. 3). The observation-based results are further corroborated through an analysis of 40 CMIP3³⁹ and CMIP5⁴⁰ climate models (see Extended Data Fig. 5 for the specific models). The 40 models, each of 186 years in record (inclusive of the past and future simulations), provide a large sample of ENSO events that is about 180 times larger than the observed sample of 25 events. Thus, the models, with their different mean climate states, provide a rigorous test bed for the effect of the current, which, along with the implications for the future, are discussed below.

The past and future climate simulations respectively correspond to the twentieth-century (1907–1999) and future projection scenarios (2006–2098) based on the Special Report on Emissions Scenarios (SRES) A1B for CMIP3 and representative concentration pathways (RCP) 4.5 for CMIP5^{39,40}. The time spans were necessarily chosen to include as many models as possible that cover the longest record without any missing data.

Mean currents and ENSO propagation structures across models. On the basis of the findings of early theoretical studies^{17,18} the prevalent direction of the basin-scale ENSO SST anomaly propagation along the Equator is an indicator for the dominant dynamical process over a given epoch: net eastwards for thermocline feedback and net westwards for zonal advective/Ekman feedback. This definition for the dominant ENSO dynamics has been adopted by previous studies^{41–43}, which we refer to hereafter as the 'ENSO propagation structure' (rather than 'ENSO mode') to tie in with the topic of our study (zonal propagation).

The diagnosis for ENSO propagation structure in observations and models has been achieved previously through a lead-lag correlation between the Niño3 index and an east-minus-west SST index which is taken as the difference between the Niño4, representing SST variability in the Central Pacific, and the Niño1+2 for the far eastern Pacific^{41,44}. The former is bounded in the west at 160°E and the latter in the east at 80°W , which is the exact longitudinal extent adopted in our study for calculating the phase transition slopes.

Here we diagnose the propagation structure in each past and future period in each model (Extended Data Fig. 5a) by the proportion of westward events (assigned as negative proportion) and eastward events (positive proportion) identified as El Niño and La Niña events with a statistically significant β . For each given period, the proportions of those four types of propagating events (that is, westward El Niño and La Niña, and eastward El Niño and La Niña; the red/blue bars and lines in Extended Data Fig. 5a) and non-propagating events (non-statistically significant slopes) add up to 1, and so the net propagation structure (grey circles for 1907–1999; black triangles for 2006–2098) can range from -1 , if all of the events propagate westwards, to $+1$ if all propagate eastwards. For example, the past ENSO events in model number 3 consist of 10% westward El Niño events, 17% westward La Niña events, 28% eastward El Niño events, 19% eastward La Niña events, and 26% non-propagating El Niño and La Niña events. Summing the proportions of the propagating events and considering the directions $(-0.1) + (-0.17) + 0.28 + 0.19$ yields an eastward propagation structure with a relative scale of 0.2 as marked by the grey circle. Although our approach is different to the commonly used correlation-based methods^{3–5}, in that we utilize β , the results using the two methods are largely consistent (figures not shown).

We find a significant positive inter-model correlation between ENSO propagation structure and mean equatorial currents (Extended Data Fig. 5b): models with weaker mean currents tend to generate a higher proportion of eastward-propagating ENSO events, and the tendency is statistically significant. This suggests that models with weak (or strong) mean currents tend to be more (or less) favourable for the thermocline feedback resulting in an eastward propagation structure (as explained in Fig. 3). Some of the models that simulate too many eastward-propagating La Niña events (Extended Data Fig. 5a; for example, models number 2, 3, 10, 17, 24 and 25), in contrast to observations (but consistent with linear theories), tend to have a weak mean current. Because the inter-model correlation between the propagation structures and mean zonal wind stresses is basically zero (Extended Data Fig. 5c), such an effect is evidence for the direct influence of the ocean currents (for example, related to specifications of the ocean model components), rather than, for instance, an effect of ENSO rectification onto the mean climate. These inter-model relationships also hold for the future simulations (see Extended Data Fig. 5 legend). Although this result has an important implication for ENSO modelling, this in itself is evidence that the ocean current does have an influence on ENSO zonal phase propagation, that is, a weaker mean current is more favourable for eastward propagation.

The model ensemble results in Extended Data Fig. 5b also imply that in a climate state with a weak background current, natural variability alone (within which the system supports naturally varying ENSO propagation structure) would more easily produce eastward-propagating events. With even weaker currents projected for the future, consistent with the weaker trade winds, the thermocline feedback effect for inducing eastward propagation is favoured further (Extended Data Fig. 6).

Previous ENSO stability analysis for a number of the CMIP3 models⁴⁵ demonstrated that the three main positive feedback processes are projected to increase, and would thus have competing effects on zonal phase propagation. The clear increase in the occurrences of eastward propagation events (Fig. 4a) can be more simply explained in terms of a weakened current as described in our study.

Effect of current reversals and models selection. One characteristic of the ENSO system is that the equatorial Pacific current anomaly is correlated with SST anomalies in the east (represented by the Niño3 index) in which the current leads Niño3 by about three months (Extended Data Fig. 4b). This highlights the tendency for an eastward (or westward) current anomaly in boreal autumn (September–December) to precede the peak of El Niño (or La Niña) in boreal winter (December–February). A particularly strong eastward anomalous current was observed during the 1982–83 and 1997–98 extreme El Niño events that leads to a re-intensified reversal in boreal autumn, a feature not seen in other events (Extended Data Fig. 4a). These extreme events are identified by their prominent eastward propagation with phase transition slope β that is stronger than in other events (Extended Data Fig. 4d). Here we demonstrate, using an aggregate of models, that current reversals have the effect of making eastward propagation characteristic more prominent. Because zonal propagation is the focus of our study, and given the dynamical links of the aforementioned features, we first select the models based on the following criteria: (1) The models must be able to simulate at least one prominent eastward-propagating El Niño event in either past or future simulation. Such an event is defined as one for which β is positive, greater than the linear-regression standard error, and above 0.5 standard deviations of all El Niño slopes (that is, following the definition for the observed slope prominence; Extended Data Fig. 4d). (2) The models must be able to simulate at least one current reversal during boreal autumn in either past or future simulation. (3) The models must produce a positive correlation between Niño3 and the current during any propagating El Niño events, a relationship also seen in observations (Extended Data Fig. 4c).

These criteria result in 24 models that simulate more realistic and distinctive current evolution between strong and moderate El Niño years (Extended Data Fig. 8) as expected from observations (Extended Data Fig. 4a), in contrast to that in the discarded models (Extended Data Fig. 9).

The effect of current reversal on zonal phase propagation is clearly exhibited by this aggregate of models, that is, it favours eastward propagation. This is due to the fact that the corresponding β tends to be more positive whenever the events coincide with a current reversal (Extended Data Fig. 7b). In the case where current reversals coincide with westward propagation, the westward slopes are found to be substantially weaker. Such an effect renders a positive correlation between the total current and β (Extended Data Fig. 7a), which is a characteristic also seen in observations (Extended Data Fig. 4d). This positive correlation further highlights the crucial role of the equatorial Pacific current on zonal phase propagation.

The effect of the total current on El Niño and La Niña propagation asymmetry is also reproduced (Extended Data Fig. 7c). The asymmetry becomes apparent with strong El Niño events, and more so when these co-occur with current reversals, similar to the observed counterpart (Fig. 2a, b; Supplementary Table 4).

An additional criterion is applied, resulting in a further exclusion of five models. Each of these excluded models already simulates 11 to 14 El Niño events with prominent eastward-propagation slope over the 93 years in the past simulation (Extended Data Fig. 10a). These are too frequent relative to the two events over the 53 years of the observational record, which translates to slightly less than four events for the 93 model years. The remaining 19 models simulate from 0 to 8 events (that is, double the expected observed frequency) in the past period. These 19 models also have climatological states that roam the regime of westward-propagation structure similar to the observed, as opposed to the five excluded models that tend to cluster about the eastward regime with already weak mean currents (Extended Data Fig. 5b; model numbers 8, 17, 24, 25 and 29). Given the extreme rarity, and to test whether a change in model climatological state can induce increased occurrence of such events, we retain the 19 models for future projections (Fig. 4).

A parallel between future projection and the late twentieth century. With the mean westward currents projected to weaken in the future (Extended Data Fig. 6), thus providing a more conducive condition for increased occurrences of current reversals (Extended Data Fig. 7d), it is expected that there will be more El Niño events exhibiting a prominent eastward-propagation characteristic in the future.

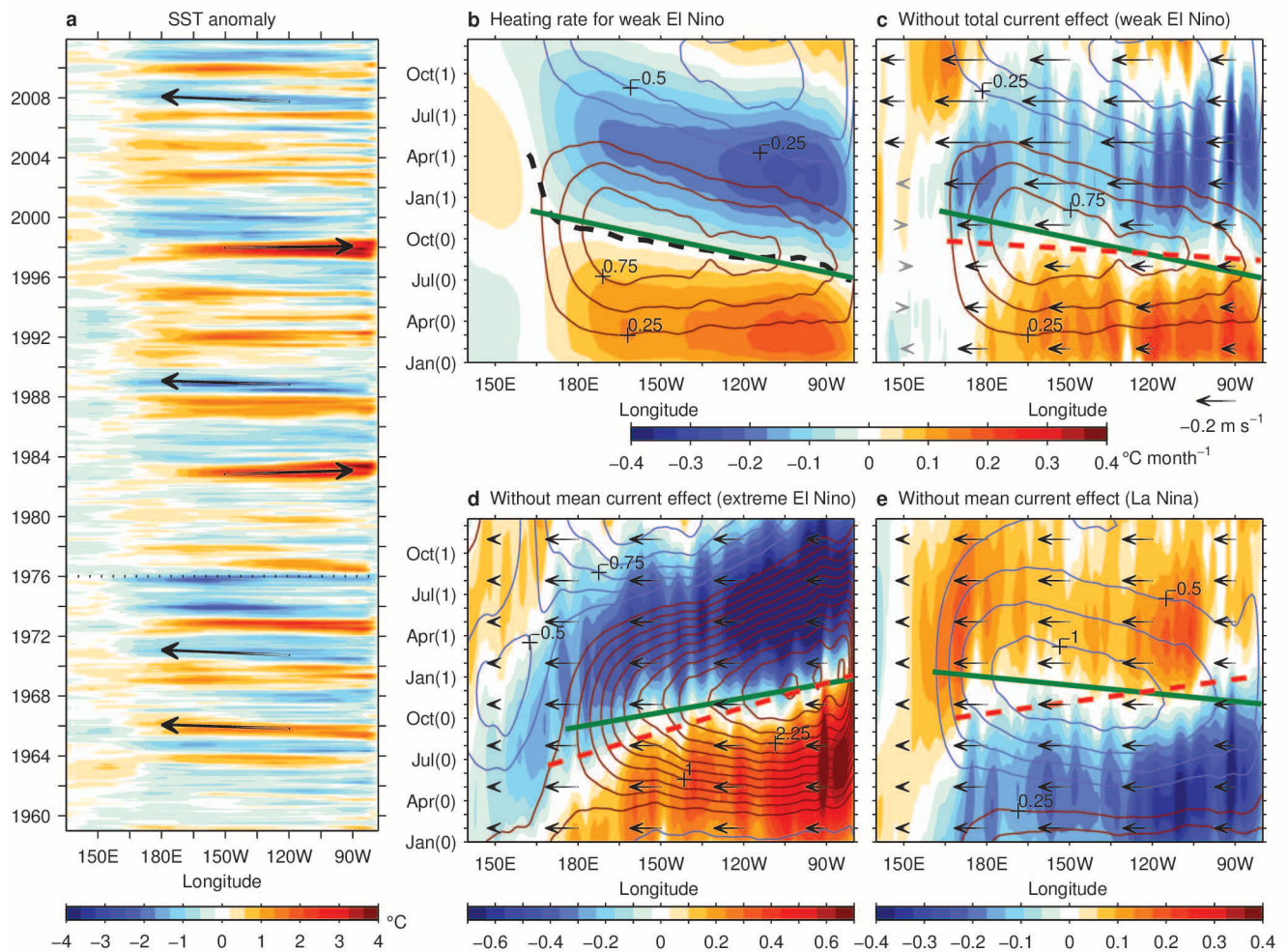
A 100% increase in the mean occurrence of such events is found (Fig. 4a), with 16 out of 19 models projecting an increase. Considering only models that simulate fewer than 8 events increases this to more than 116%, with model consensus consistently above 83%. As expected, retaining those that already simulate frequently

occurring events (that is, saturated with eastward propagation) reduces the amount of increase to 76% when including models that already simulate up to 11 events, and 46% using all of the 24 models. Nonetheless, in all cases, the models as an aggregate simulate a notable increase in future occurrences of eastward-propagating El Niño events that is significant well above the 95% confidence level, with at least 75% of the models projecting an increase.

As revealed by the observational analysis, the emergence of an eastward propagation in the post-76 period is in part because the mean westward current is weaker, and in part because the eastward current anomalies associated with the extreme El Niño events are sufficiently large to reverse the current. In this regard, the variety of events and mean states provided by the different models point to a slightly different scenario for the future in which the importance of the projected current weakening is highlighted. This is evident as the model consensus is weak in the projection for stronger ENSO amplitude (11 out of 19 models; Extended Data Fig. 10c), but all of the models project a weaker mean current (Extended Data Fig. 10b). Despite this, there is still a tendency for stronger increase in the number of eastward-propagating events in models that also project a larger increase in ENSO amplitude (Extended Data Fig. 10c). This is through the contribution by current reversals (Extended Data Fig. 10d), which tend to occur with stronger El Niño events and have the effect of making the eastward propagation characteristic more prominent (Extended Data Fig. 7).

We note that although a weaker mean current facilitates current reversal, such that any modest eastward current anomaly can more easily exceed the background current, the increase in the number of current reversals in the future (Extended Data Fig. 7d) does not always translate to more occurrences in events having a prominent eastward-propagation characteristic (Extended Data Fig. 10d). This is expected, given the various kinds of event concurrences that the model aggregate provides (Extended Data Fig. 7). In fact, although all of the increase in eastward-propagating events is associated with weaker mean currents and El Niño events of all magnitudes (Fig. 4b and c), only 45% of this is associated with current reversal events, within which 85% are associated with large-magnitude El Niño events (Fig. 4d). Thus, given the weak model consensus in projecting an increase in ENSO amplitude, the most robust feature shared between the future projection and the change observed during the late twentieth century is the weaker westward mean current, which is projected by all of the models.

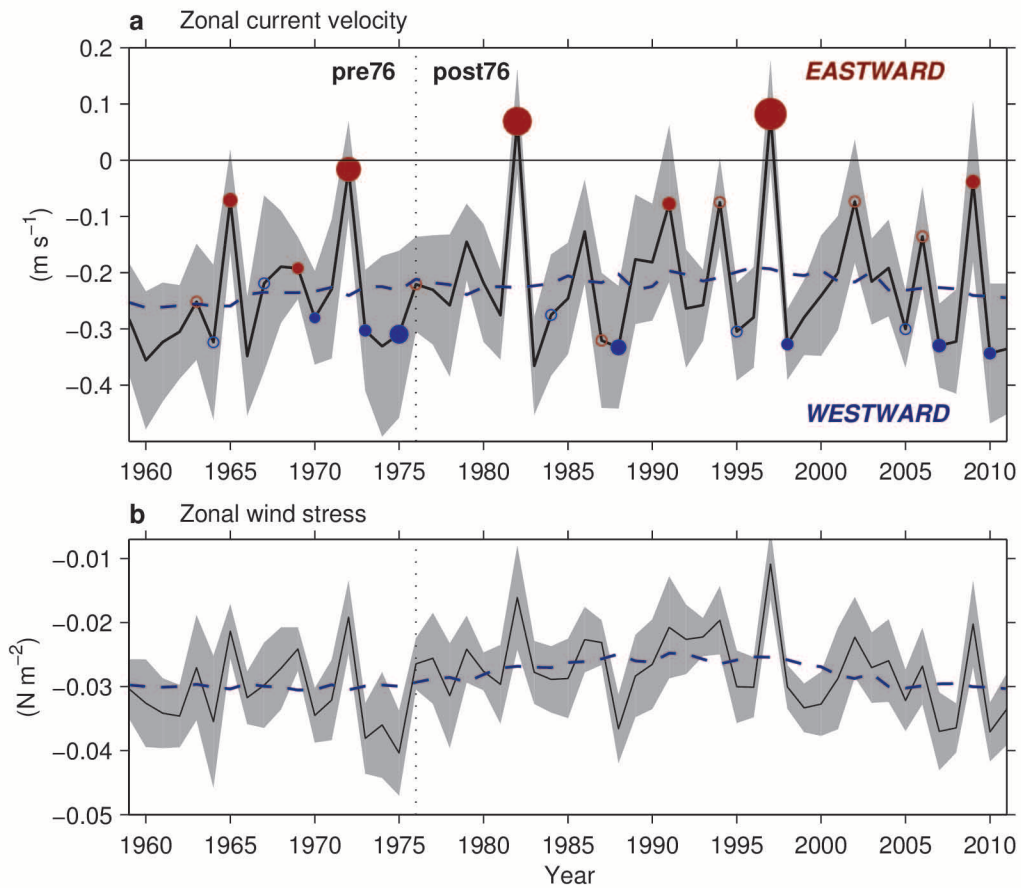
- Balmaseda, M. A., Vidard, A. & Anderson, D. The ECMWF ocean analysis system: ORA-S3. *Mon. Weath. Rev.* **136**, 3018–3034 (2008).
- Balmaseda, M. A., Mogenssen, K. & Weaver, A. Evaluation of the ECMWF Ocean Reanalysis ORAS4. *Q. J. R. Meteorol. Soc.* **139**, 1132–1161 (2013).
- Carton, J. A. & Giese, B. S. A reanalysis of ocean climate using simple ocean data assimilation (SODA). *Mon. Weath. Rev.* **136**, 2999–3017 (2008).
- Giese, B. S. & Ray, S. El Niño variability in simple ocean data assimilation (SODA), 1871–2008. *J. Geophys. Res.* **116**, C02024 (2011).
- Behringer, D. W. The Global Ocean Data Assimilation System at NCEP. In *11th Symp. on 'Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface'* 1–12 (AMS 87th Annual Meeting, Henry B. Gonzales Convention Center, 2007).
- Roberts, J. & Roberts, T. D. Use of the Butterworth low-pass filter for oceanographic data. *J. Geophys. Res.* **83** (C11), 5510–5514 (1978).
- Smith, T. M., Reynolds, R. W., Peterson, T. C. & Lawrimore, J. Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Clim.* **21**, 2283–2296 (2008).
- Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* 168–232, Chs 13–16 (Chapman & Hall, 1993).
- Meehl, G. A. et al. The WCRP CMIP3 multimodel dataset: a new era in climate change research. *Bull. Am. Meteorol. Soc.* **88**, 1383–1394 (2007).
- Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- Guilyardi, E. El Niño-mean state-seasonal cycle interactions in a multi-model ensemble. *Clim. Dyn.* **26**, 329–348 (2006).
- Santoso, A., England, M. & Cai, W. Impact of Indo-Pacific feedback interactions on ENSO dynamics diagnosed using ensemble climate simulations. *J. Clim.* **25**, 7743–7763 (2012).
- Aiken, C. M., Santoso, A., McGregor, S. & England, M. H. The 1970's shift in ENSO dynamics: a linear inverse model perspective. *Geophys. Res. Lett.* **40**, 1612–1617 (2013).
- Trenberth, K. E. & Stepaniak, D. P. Indices of El Niño evolution. *J. Clim.* **14**, 1697–1701 (2001).
- Kim, S.-T. & Jin, F.-F. An ENSO stability analysis. Part II: results from the twentieth and twenty-first century simulations of the CMIP3 models. *Clim. Dyn.* **36**, 1609–1627 (2011).



Extended Data Figure 1 | Zonal propagation of SST anomalies and effect of current on mixed layer heat balance during ENSO events. **a**, SST³⁷ anomalies along the equatorial Pacific (averaged between 5° S–5° N) over January 1959 to December 2011, with seasonal cycle and linear trend (referenced to the entire 1959–2011) removed. The arrows, whose slopes are calculated from the multi-reanalysis ensemble average, indicate zonal propagation directions.

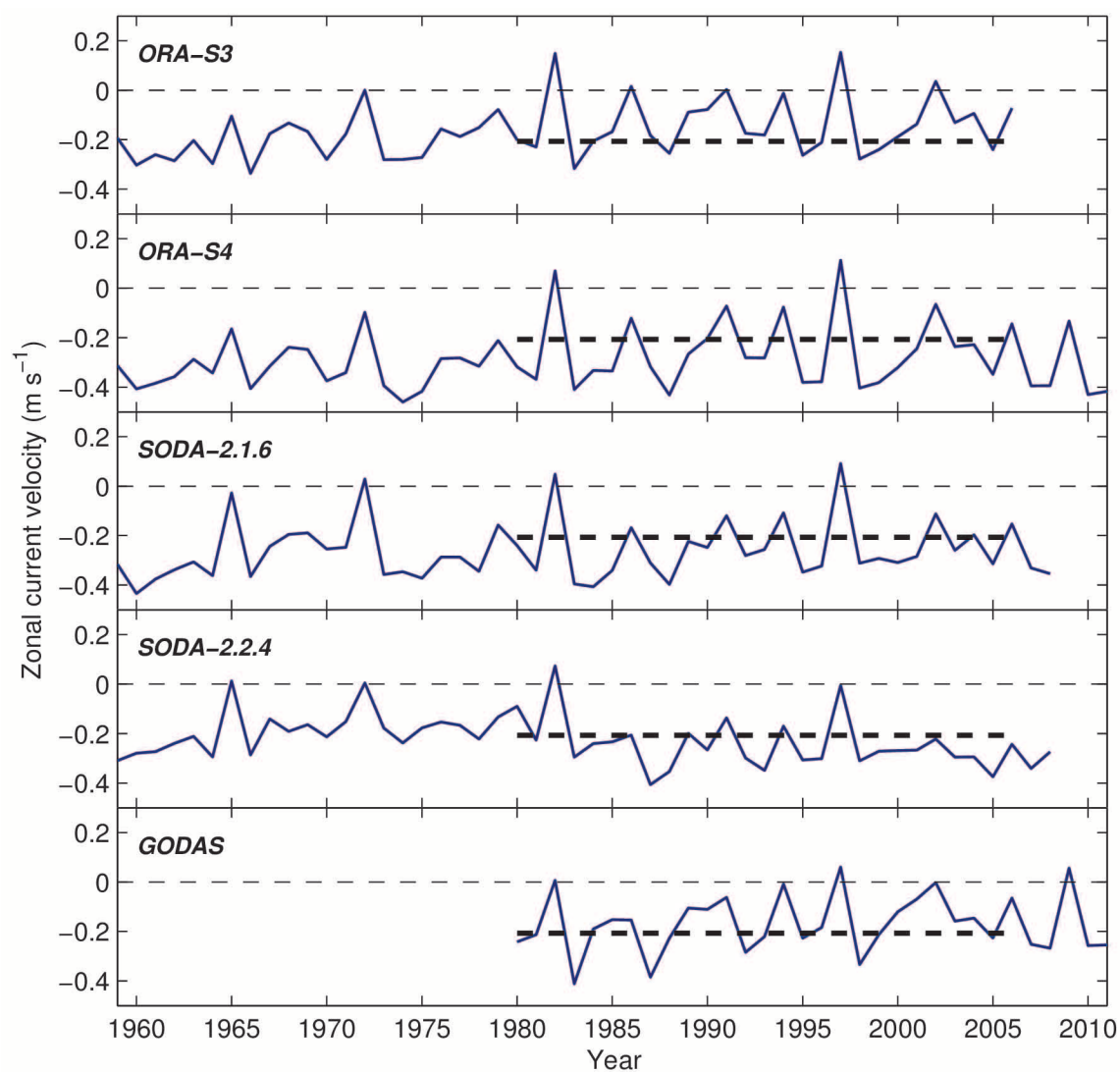
b, Composite evolution of interannual-scale heating rate (colour scale; in units of °C per month) of the equatorial Pacific mixed layer during post-1976 moderate El Niño events. The phase transition (dashed black line) tracks the evolving peak of temperature anomaly (red contours are positive and blue

contours are negative) with a statistically significant linear fit slope (green line; $\beta = -0.97$, $P < 0.01$). **c**, As in **b** but with advection due to the total current (arrow) removed, resulting in a $\beta^* = -0.29$ (red dashed line) value that is statistically significant ($P < 0.05$). Only statistically significant values above the 95% confidence level are shaded in colour, contoured, or marked by black arrows (grey arrows otherwise). **d**, As in **c** but for extreme El Niño events (1982, 1997) with the effect of mean current (arrows) removed. **e**, As in **d** but for post-1976 La Niña events. The β^* values are 1.44 in **d** and 0.61 in **e** and are statistically significant ($P < 0.01$).



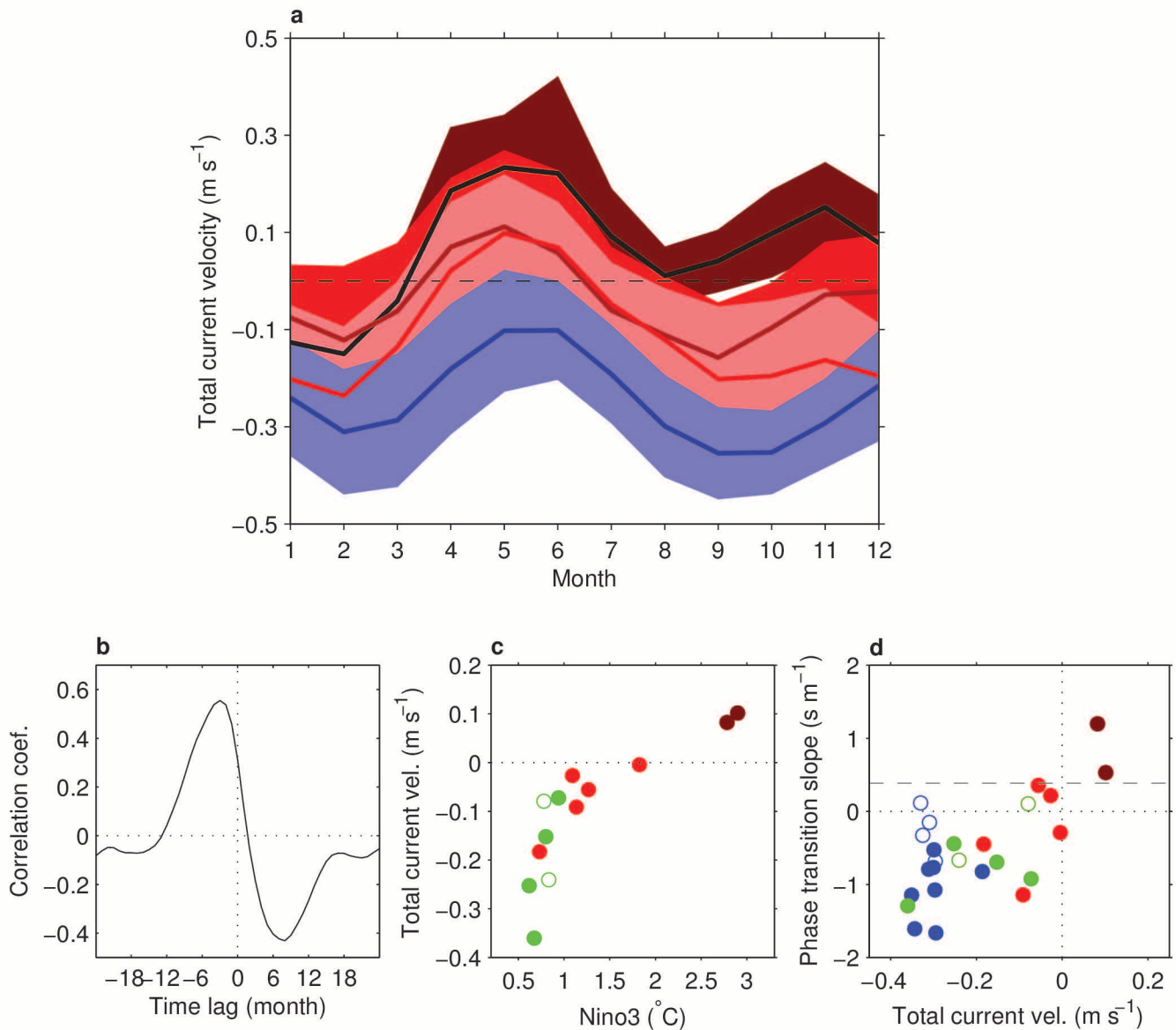
Extended Data Figure 2 | Time evolution of equatorial Pacific zonal current and wind stress. **a**, The same as Fig. 1c for zonal current velocity averaged across the reanalysis products. The dashed curve highlights interdecadal

variation using a 13-year running mean. Grey shading denotes two standard deviation about each mean value, representing monthly spread and variations across reanalyses. **b**, As in **a** but for surface zonal wind stress.



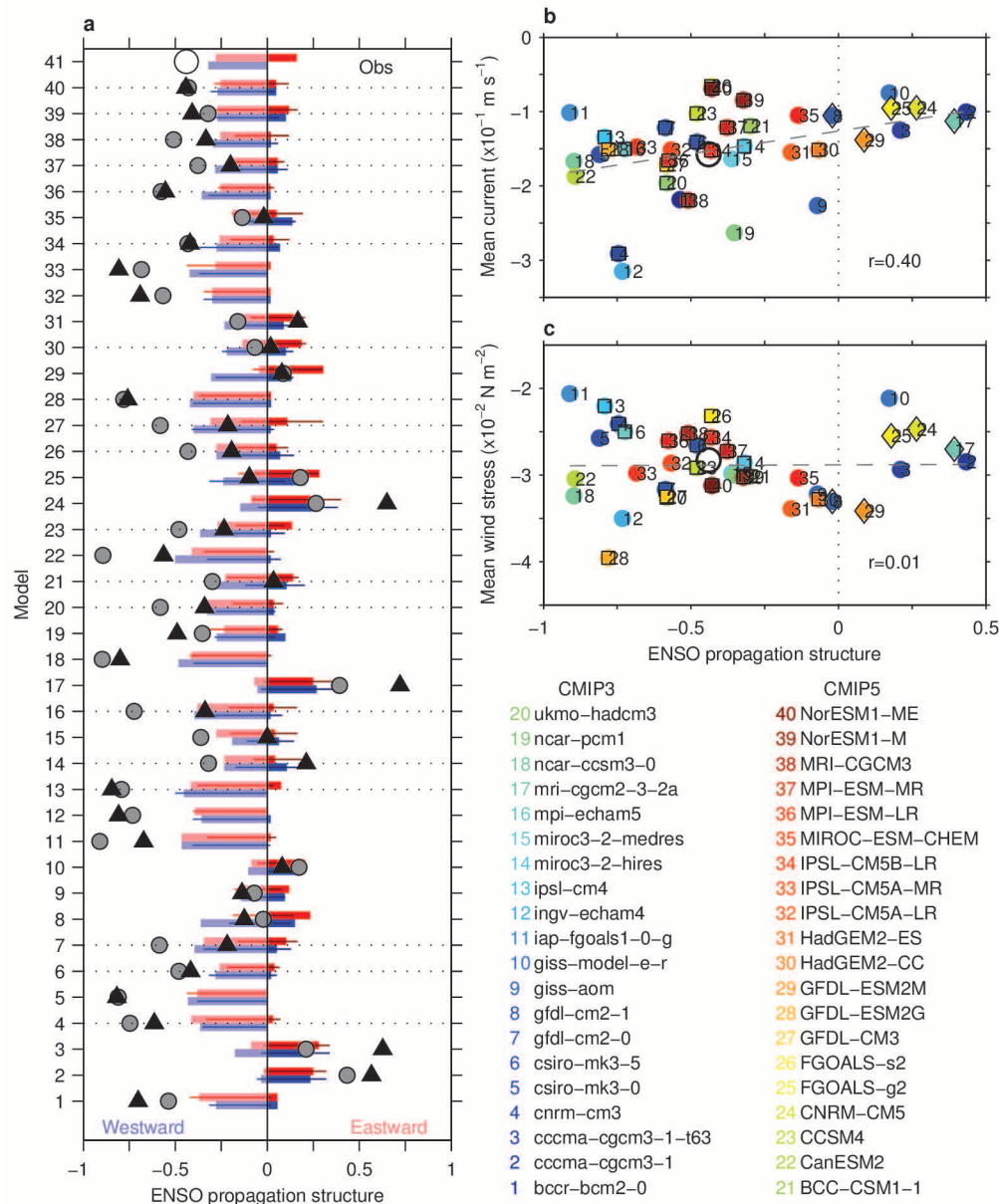
Extended Data Figure 3 | Time evolution of the equatorial Pacific zonal current across reanalysis products. Raw time series of zonal current velocity averaged over 5° S–5° N, 160° E–90° W, capturing the Niño4 to Niño3 regions,

and over the ENSO development phase (August to December). The ensemble average (1980–2006) is marked by the thick horizontal dashed line.



Extended Data Figure 4 | Observed characteristics of equatorial Pacific current associated with ENSO. **a**, Total current evolution composited over developing phase of ENSO: extreme El Niño (dark red shading/black line), strong El Niño (red shading/dark red line), weak El Niño (pink shading/red line), and La Niña (blue shading/dark blue line). Thick lines indicate the mean composites, and the coloured shades are for one standard deviation unit above and below the means representing the spread across the different reanalyses and each classified events. **b**, Lead-lag monthly correlation between the reanalysis ensemble average current and Niño3 with eastward current anomalies leading warm Niño3 anomalies at three months. **c**, Total current

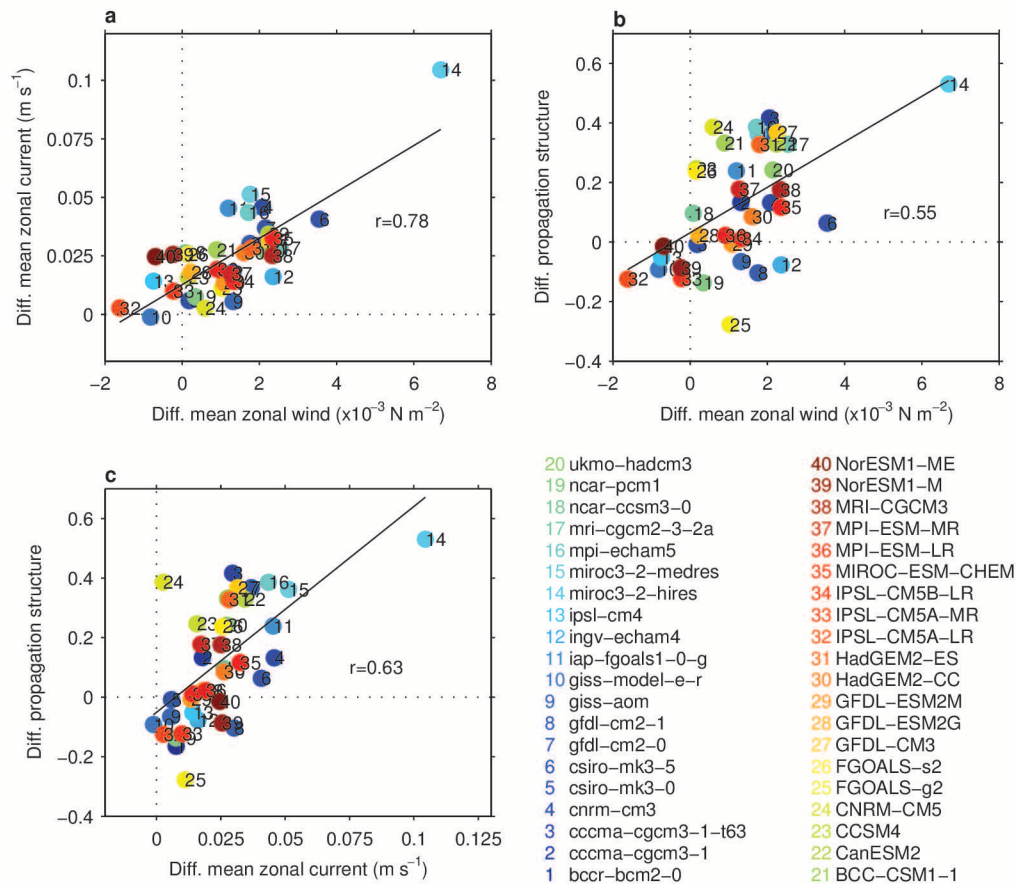
velocity (vel.) averaged over September–December versus Niño3 (December–February) associated with extreme (dark red circles), strong (red circles), and weak (green circles) El Niño events in each pre-1976 and post-1976 period, with a correlation coefficient of 0.82, significant at the 99% level. Open circles indicate non-statistically significant β . The correlation ($r = 0.84$) remains significant at the 99% level even when these points are excluded. **d**, As in **c** but for total current versus β during all ENSO events (blue circles for La Niña). The correlation coefficient (coef.) between current and statistically significant β for El Niño is $r = 0.75$, which is significant at the 99% level. The dashed horizontal line in **d** marks half a standard deviation unit of all the El Niño slopes.



Extended Data Figure 5 | ENSO propagation structure in CMIP models.

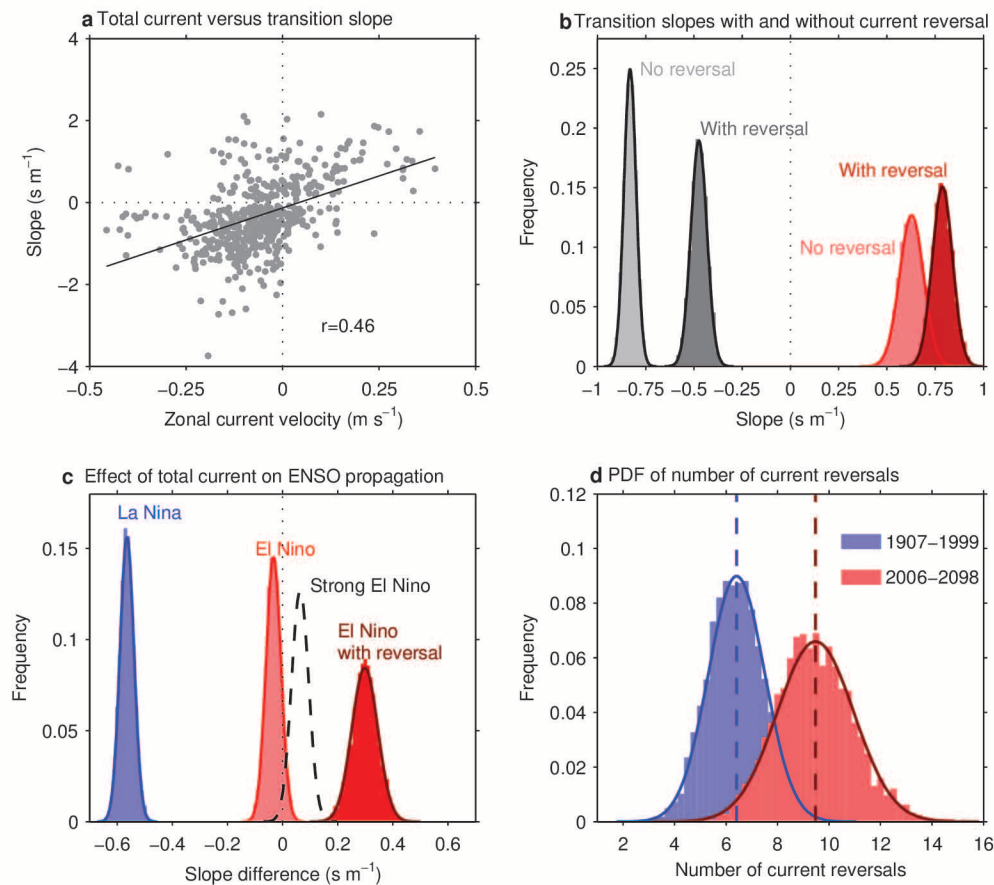
a, Propagation structure in each CMIP model (grey circles for 1907–1999; black triangles for 2006–2008 period) and observations (1959–2011; the large open circle labelled ‘Obs.’). The propagation structure is defined by summing up the proportion of westward events (negative proportion) and eastward events (positive proportion) identified as El Niño (red bar for 1907–1999; red line for 2006–2008) and La Niña (blue bar for 1907–1999; blue line for 2006–2008) with statistically significant β . The colour intensities for the bars and lines indicate the four types of propagating events. The proportions of propagating events and non-propagating events add up to 1, and so the net propagation structure (grey circle or black triangle) can range from a scale of -1 if all events propagate westwards to $+1$ if all propagate eastwards. Eastward (westward)

propagation structure is an indication for a more dominant thermocline (zonal advective) feedback mechanism. **b**, Propagation structure versus long-term annually averaged zonal current velocity across all CMIP models (coloured markers) in the past simulation, revealing a positive correlation ($r = 0.40$) significant at the 95% level ($r = 0.44$ for future). Open circle marks the observed counterpart using data from 1959 to 2011 for a larger event sample. **c**, As in **b** but for mean zonal wind stress, exhibiting no significant correlation ($r = 0.01$; $r = 0.14$ for the future). Models marked by dotted horizontal lines in **a** and squares in **b** and **c** indicate those selected for future projections (Fig. 4). Models marked with diamonds in **b** and **c** simulate realistic flow features but are saturated with eastward-propagating events that they have already produced in the past simulation (see Extended Data Fig. 10a).



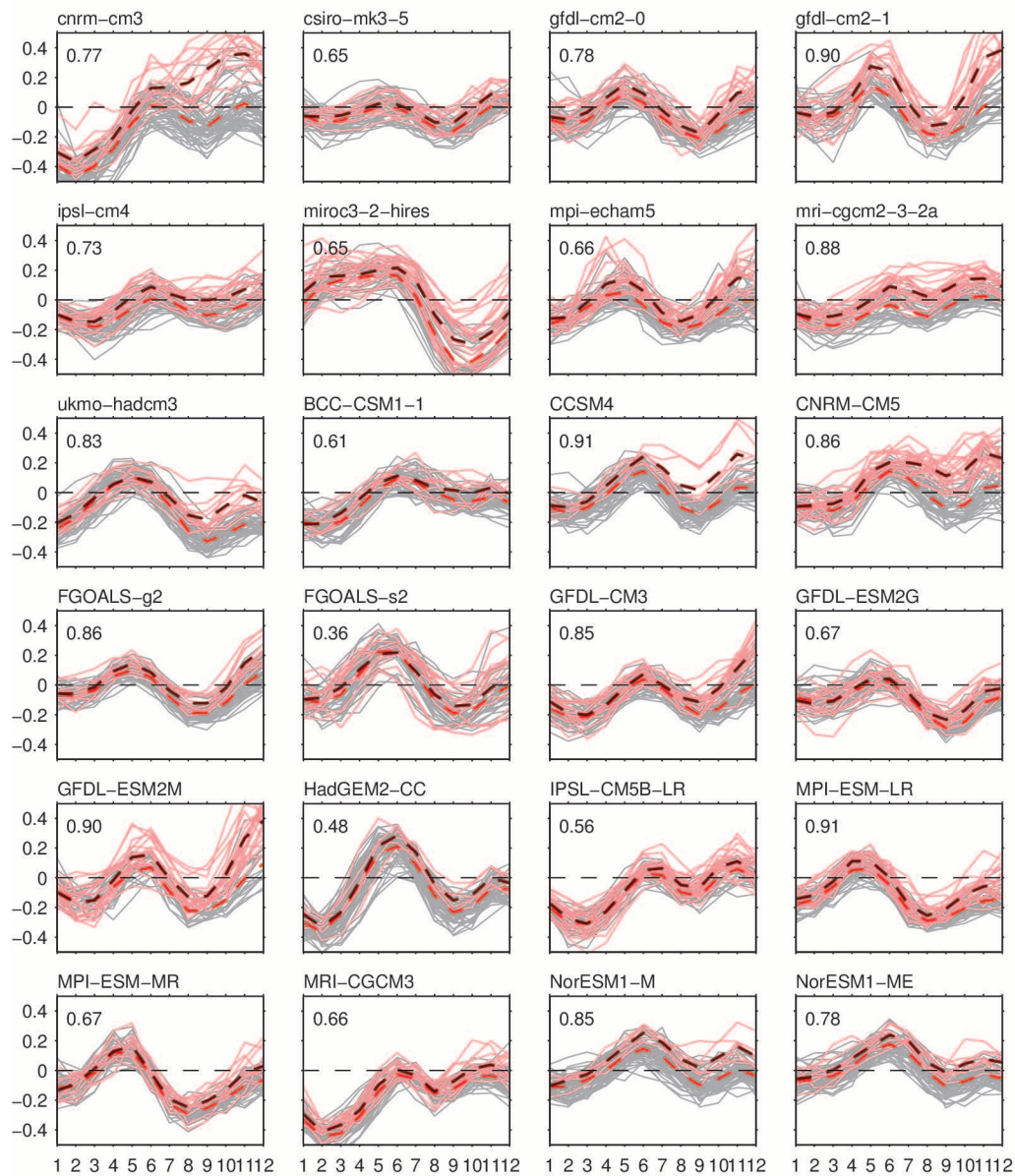
Extended Data Figure 6 | Projected changes of long-term mean zonal wind stress, zonal current velocity, and propagation structure across the CMIP3 and CMIP5 models. **a**, Future and past difference (diff.) in long-term mean zonal wind stress and zonal current velocity. **b**, Future and past difference in long-term mean zonal wind stress and ENSO propagation structure (Extended Data Fig. 5). **c**, Future and past difference in long-term mean zonal current

velocity and ENSO propagation structure. The correlations between each of the variables are shown in the panels and are statistically significant at the 99% level. Removing the model outlier (miroc3-2-hires) reduces correlations in **a**, **b** and **c** to 0.61, 0.47 and 0.59, respectively, but are still statistically significant up to the 99% level.



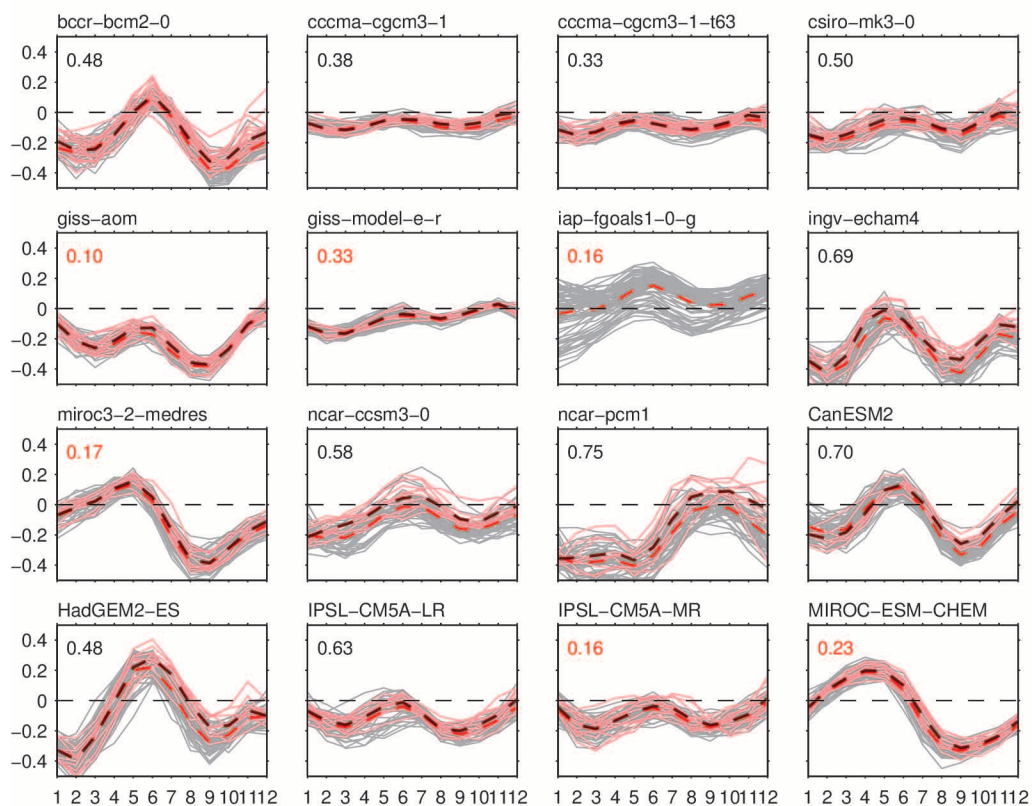
Extended Data Figure 7 | Effect of current reversals on zonal phase propagation and future projection. The analysis incorporates the 24 models that simulate realistic flow features (see Methods). **a**, Correlation between total current and phase transition slope during El Niño events in the past simulation (1907–1999). The positive correlation ($r = 0.46$), significant above the 99% level (with 472 data points), confirms the relationship seen in the limited observational record (Extended Data Fig. 4). **b**, Probability density of β for westward (grey) and eastward (red) El Niño events with (darker shading) and without (lighter shading) current reversals. **c**, Probability density of the difference in phase transition slope before and after the effect of total current

removed from the heat balance ($\beta - \beta^*$), for all La Niña events (blue), all El Niño events (light red), and El Niño events that co-occur with current reversals (darker red). The probability density for strong El Niño events (greater than one standard deviation) is shown by the dashed curve. **d**, Probability density of number of current reversals associated with any events in the past (1907–1999; blue) and future (2006–2098; red) periods. Vertical lines in **d** indicate the respective mean values (6.4 and 9.5 for past and future periods, respectively). The statistics in **b**, **c** and **d** are generated using a bootstrap sampling technique with 5,000 simulations.

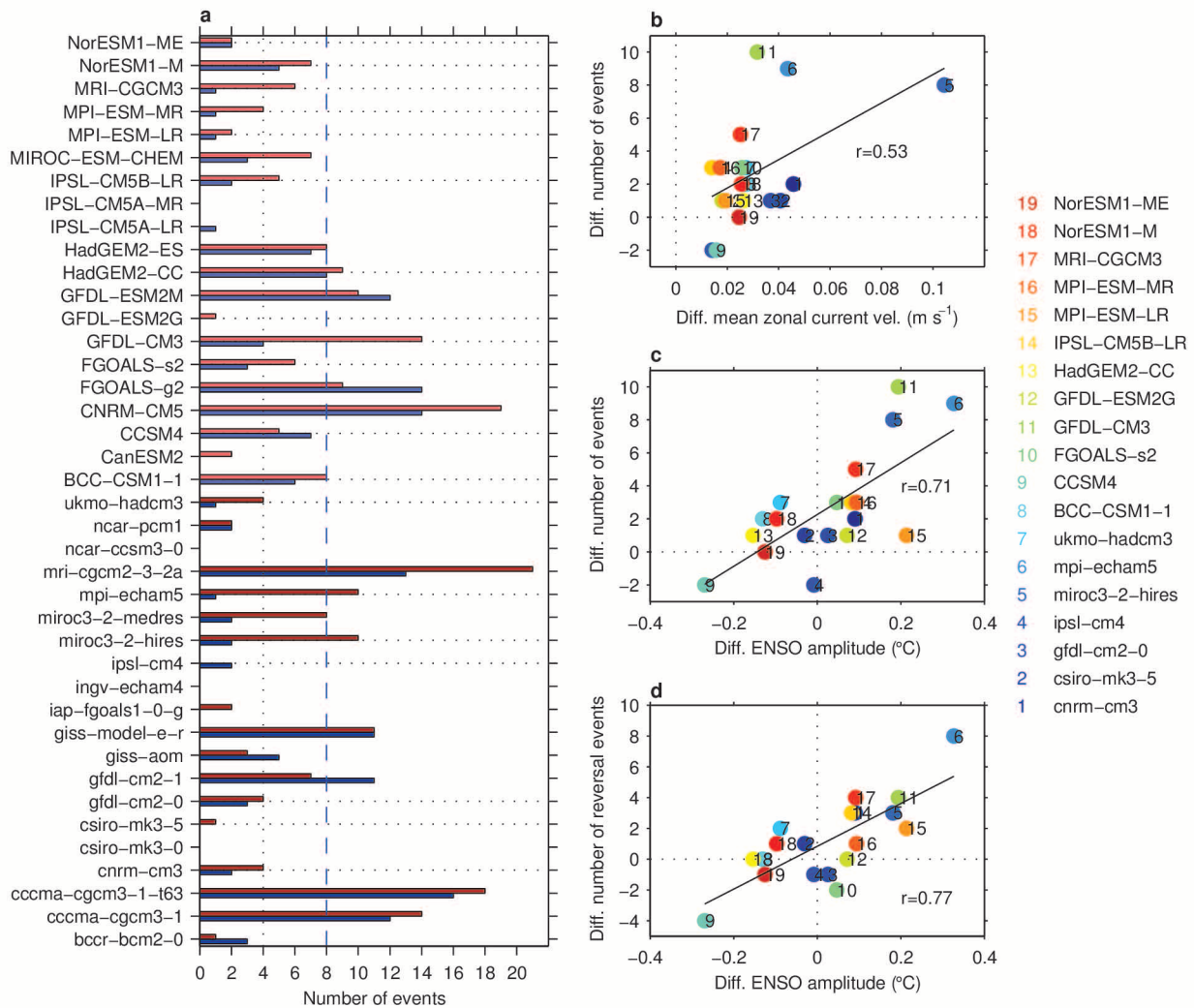


Extended Data Figure 8 | Monthly evolution of the total current during the developing year of El Niño events in the selected CMIP3 and CMIP5 models. Red and grey curves respectively represent El Niño events in both past and future simulations that are classified as above and below 1.5 standard deviations of Niño3 (December–February average), normalized by the standard deviation

of the past period. Only events with statistically significant transition slopes are considered. The corresponding dashed curves indicate the sample averages. Each panel displays the correlation coefficient between the equatorial Pacific current (September–December average) and the Niño3 anomalies, following the observed counterpart (Extended Data Fig. 4c).



Extended Data Figure 9 | As Extended Data Fig. 8, but for the excluded models. Correlation coefficients displayed in red are not statistically significant.



Extended Data Figure 10 | Occurrences of El Niño with prominent eastward propagation and future projection as a function of mean current, ENSO amplitude, and current reversals in the CMIP3 and CMIP5 models.

a, Number of events for each of the 40 models for the past (1907–1999; blue) and future (2006–2098; red) periods (see Methods for event criteria). The number of events over the 93 model years expected from the observed occurrences is four (dotted vertical line). For future projection, we consider models that produce an occurrence of 0–8 events (that is, doubling; dashed

vertical line). Dotted horizontal lines indicate the selected models. **b**, Future and past difference (diff.) in event occurrences against that of the long-term mean zonal current velocity (vel.). **c**, As in **b** but against the future and past difference in ENSO amplitude as defined by the standard deviation of Niño3 index. **d**, ENSO amplitude difference against the difference in number of eastward-propagating events with current reversals. The correlation coefficients displayed in the panels are significant at the 95% level.

Foundering of lower island-arc crust as an explanation for the origin of the continental Moho

Oliver Jagoutz¹ & Mark D. Behn²

A long-standing theory for the genesis of continental crust is that it is formed in subduction zones¹. However, the observed seismic properties of lower crust and upper mantle in oceanic island arcs^{2,3} differ significantly from those in the continental crust⁴. Accordingly, significant modifications of lower arc crust must occur, if continental crust is indeed formed from island arcs. Here we investigate how the seismic characteristics of arc crust are transformed into those of the continental crust by calculating the density and seismic structure of two exposed sections of island arc (Kohistan and Talkeetna). The Kohistan crustal section is negatively buoyant with respect to the underlying depleted upper mantle at depths exceeding 40 kilometres and is characterized by a steady increase in seismic velocity similar to that observed in active arcs. In contrast, the lower Talkeetna crust is density sorted, preserving only relicts (about ten to a hundred metres thick) of rock with density exceeding that of the underlying mantle. Specifically, the foundering of the lower Talkeetna crust resulted in the replacement of dense mafic and ultramafic cumulates by residual upper mantle, producing a sharp seismic discontinuity at depths of around 38 to 42 kilometres, characteristic of the continental Mohorovičić discontinuity (the Moho). Dynamic calculations indicate that foundering is an episodic process that occurs in most arcs with a periodicity of half a million to five million years. Moreover, because foundering will continue after arc magmatism ceases, this process ultimately results in the formation of the continental Moho.

Continental crust is characterized by a lower-velocity upper crust (seismic P-wave velocity $V_p \approx 5\text{--}6\text{ km s}^{-1}$) and a higher-velocity lower

crust ($V_p \approx 7\text{--}7.5\text{ km s}^{-1}$), separated from the underlying mantle ($V_p \approx 8\text{--}8.5\text{ km s}^{-1}$) by the sharp Moho⁴. Globally, with the exception of active orogenic belts or rifts, the continental Moho occurs at a relatively constant depth of $41 \pm 6\text{ km}$ (ref. 4). In contrast, the seismic structure of the lower crust in many active arcs is defined by a transitional increase from lower crustal velocities of $V_p \approx 7\text{ km s}^{-1}$ to sub-Moho velocities of $V_p \approx 7.6\text{--}7.7\text{ km s}^{-1}$, significantly slower than the sub-Moho velocities observed in continental regions. The sharply defined Moho seen in continental crust is generally absent in arcs^{5–7}; instead a weak discontinuity (increase in V_p from about 6.8 to 7.2 km s^{-1}) is observed that has been interpreted to indicate either a contact between mafic lower crust and unusually hot upper mantle or an intra-crustal contact between mafic and ultramafic cumulates⁶. Accordingly, if continental crust is formed in arcs, significant reworking of arc lower crust must occur to transform the transitional lower-crust–mantle interface in arcs into the sharply defined crust–mantle discontinuity of continental regions.

It is widely accepted that mafic/ultramafic rocks in arc lower crust can become denser than the underlying upper mantle and could founder back into the upper mantle^{8–10}. This process can explain the andesitic chemical composition of continental crust ultimately derived from basaltic mantle melts^{10,11}. Previous studies have proposed that the maximum observed thickness of the continental crust (about $70\text{--}80\text{ km}$) is controlled by the depth interval at which a density inversion occurs¹²; however, the relationship between lower crustal foundering and the location and nature of the Moho has not been established. Specifically, crustal

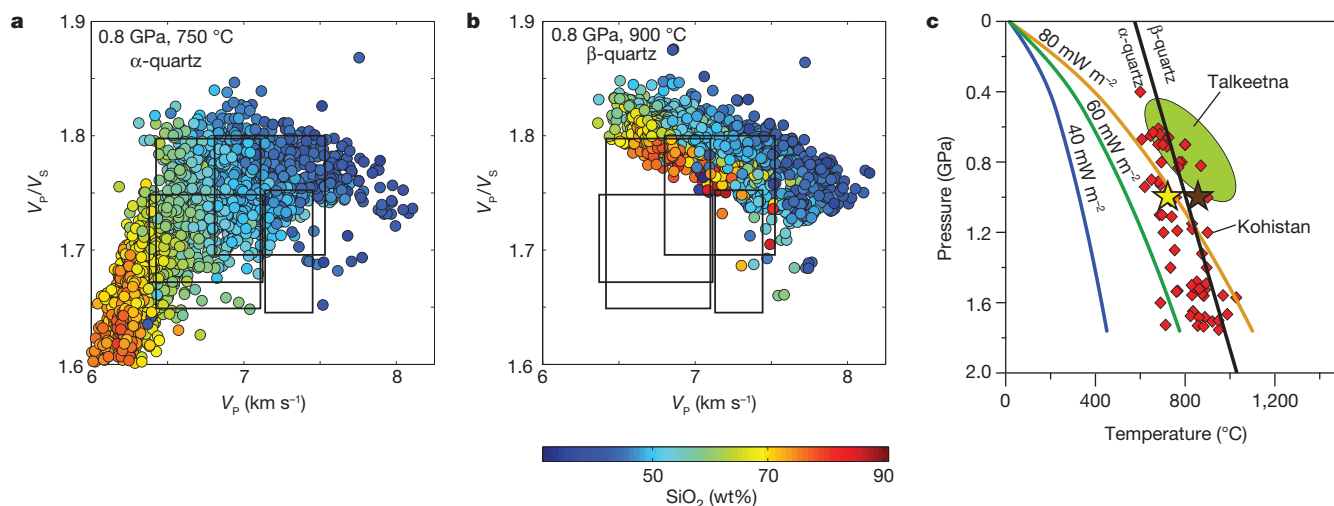


Figure 1 | Seismic and petrological constraints on the thermal regime in arcs. **a, b**, Seismic velocities of representative lower-crustal rocks from continents and arcs in the α -quartz (**a**) and β -quartz (**b**) stability fields ($n = 428$). Boxes indicate the seismic properties observed in the lower crust of active arcs (see Methods for references). **c**, Pressure versus temperature diagram showing the location of the α -quartz to β -quartz transition and

metamorphic pressure and temperature recorded in the Kohistan and Talkeetna sections^{18,19,31}. Yellow and brown stars indicate the pressure and temperature conditions used to calculate panels **a** and **b**, respectively. The observed V_p/V_s and V_p values constrain the spatially averaged temperatures to lie within the α -quartz field¹⁷.

¹Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA. ²Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

rocks can become density unstable with respect to the upper mantle over a significant depth interval from 20 to 60 km or more, depending on compositions and temperature conditions in the arc lower crust (see Methods for detailed discussion). Detailed knowledge of the composition and temperature regime in the arc lower crust is therefore essential to assess how such a foundering process could influence the seismic properties of the crust–mantle interface.

To constrain the depth at which foundering occurs in arcs we calculate the density and seismic properties of rocks from Kohistan and Talkeetna, the two best-exposed oceanic arc sections (see Methods for geological setting). Previous studies have suggested that the Talkeetna arc crust is generally less dense than the underlying upper mantle peridotites ('density stable')^{13,14}, whereas in Kohistan lower-crustal rocks denser than the upper-mantle peridotites ('density unstable') are preserved¹⁵. Here we present thermodynamic modelling of observed crustal compositions at pressures and temperatures appropriate for the formation of the arc sections (see Methods for details of the modelling). We use these results to reconstruct the detailed density and seismic structure of the two arc sections during their formation to (1) determine the depth at which the density inversion in arcs occurs, and (2) explore the effect of foundering on the seismic properties of the arc lower crust.

To calculate the seismic/density structure of the Kohistan and Talkeetna arcs, we first estimated the temperature in the lower crust of active arcs. Although the thermal structure of an active arc is transient owing to the interaction between a conductive geothermal gradient and perturbations from frequent melt infiltration events (see ref. 16 for example), we can use V_p/V_s (where V_s is shear-wave velocity) estimates from the arc lower crust in combination with geothermometry on metamorphic mineral assemblages to infer the spatially averaged thermal conditions during the construction of the arc crust. Estimates show that V_p/V_s in the lower crust of active arcs is variable, but is generally 1.70–1.80 with a corresponding V_p of 6.5–7.5 km s⁻¹ (Fig. 1a, b)¹⁷. This low V_p/V_s indicates that quartz-bearing lithologies are present in the arc lower crust and the quartz must be mostly the low-temperature alpha-quartz pseudomorph (Fig. 1a, b).

These observations constrain the spatially averaged temperature in arc lower crust to less than 800–850 °C at approximately 25–40 km depth, consistent with a conductive geothermal gradient of about 60–70 mW m⁻² (Fig. 1c). Similar metamorphic temperatures are preserved in Kohistan (700–800 °C at about 40–50 km; ref. 18), whereas higher temperatures are recorded in Talkeetna (about 900–1,000 °C at depths of around 40 km; ref. 19) (Fig. 1c). On the basis of these results we calculated density and seismic properties along appropriate geotherms for Kohistan (60 mW m⁻²) and Talkeetna (80 mW m⁻²) (Fig. 1c). However, as discussed in the Methods and shown in Extended Data Fig. 1, the effect of temperature on key metamorphic reactions controlling density and seismic structure is modest and does not influence the main conclusions of this study.

In both the Kohistan and Talkeetna sections an abrupt increase in V_p is observed between the dominant felsic/intermediate plutonic rocks of the upper crust (6.3–6.4 km s⁻¹) and underlying mafic arc crust (6.9–7.1 km s⁻¹) (Figs 2 and 3). However, the lower crust in the two arc sections differs significantly. In Kohistan, V_p in the lower crust increases linearly between 35 km depth and 50 km depth with two minor discontinuities (Figs 2 and 3). The first is an increase in V_p at about 40 km depth between gabbroic rock (about 7.0 km s⁻¹) and mafic garnet granulite (about 7.5 km s⁻¹). The second is an increase in V_p between the garnet granulite and the underlying ultramafic rocks (about 8.0 km s⁻¹) at approximately 50 km depth. This contact between garnet granulite and ultramafic rocks has traditionally been interpreted to reflect the seismic Moho of the Kohistan arc^{20,21}. The discontinuity at around 40 km coincides with a density inversion where the lowermost 10 km of crust is significantly denser ($\Delta\rho = \rho_{\text{crust}} - \rho_{\text{mantle}} = 40\text{--}280 \text{ kg m}^{-3}$) than the underlying mantle (Figs 2 and 3). This density inversion corresponds to a pressure of about 1.0–1.2 GPa and is related to the appearance of garnet as a stable phase in mafic lithologies (the 'garnet-in' reaction; see Methods). Rocks above this discontinuity are generally

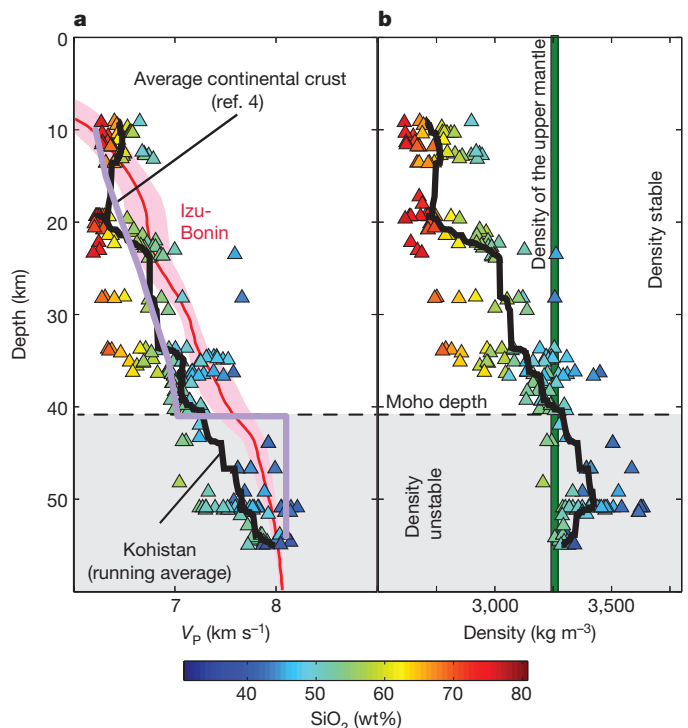


Figure 2 | Detailed V_p and density depth-structure of the exposed Kohistan arc section compared to the average continental crust and the Izu-Bonin arc crust. **a**, The seismic characteristics of the reconstructed Kohistan arc are similar to those of the Izu-Bonin arc crust (pink line, with pink shading indicating the variability)³² and differ significantly from the average seismic characteristics of continental lower crust (purple)⁴. Specifically, a sharp Moho that defines the crust–mantle interface in continents is absent in arcs. **b**, The depth of the continental Moho coincides with the depth at which crustal rocks from Kohistan become density-unstable with respect to the depleted upper mantle (green line). Black lines indicate the running average.

density-stable compared to a depleted upper mantle, whereas the rocks below are generally density-unstable and could foundering back into the upper mantle.

In Talkeetna, the Moho is a sharp contact between the basal gabbro and the underlying depleted mantle²² occurring at pressures (about 1 ± 0.14 GPa; ref. 19) comparable to those of the observed density inversion in Kohistan, and corresponding to a maximum crustal thickness of around 40 km depth (Fig. 3). Petrological considerations indicate that significant volumes of mafic/ultramafic cumulates are missing from the base of the Talkeetna arc²². Density-unstable garnet granulites ($V_p \approx 7.5\text{--}7.7 \text{ km s}^{-1}$), similar to those preserved in Kohistan, are only present as relicts in a thin layer (less than about 100 m thick) situated between the basal gabbro and the upper mantle (Fig. 2)²². With the exception of these garnet granulites, the Talkeetna arc crust is generally density-stable ($\Delta\rho = \approx -160 \text{ kg m}^{-3}$) and a single large increase in V_p is calculated at around 40 km depth, between gabbroic rocks (about 7.0 km s⁻¹) and the underlying depleted harzburgite (about 7.9–8.0 km s⁻¹).

The calculated seismic properties of the density-stable Talkeetna lower crust match those of the continental lower crust, whereas the density-unstable Kohistan lower crust has seismic characteristics comparable to the sub-Moho structure in active arcs (Fig. 3 and Extended Data Fig. 2). An important difference between the two sections is that the Talkeetna arc crust is density sorted, whereas the lower Kohistan arc is not (Fig. 3). Density sorting of the Kohistan arc lower crust, in which unstable cumulates are replaced by harzburgitic sub-arc mantle, would result in a lower crust with seismic properties comparable to those of Talkeetna and the continental lower crust (Figs 2 and 3). From these observations we propose that density sorting of arc lower crust is a crucial mechanism

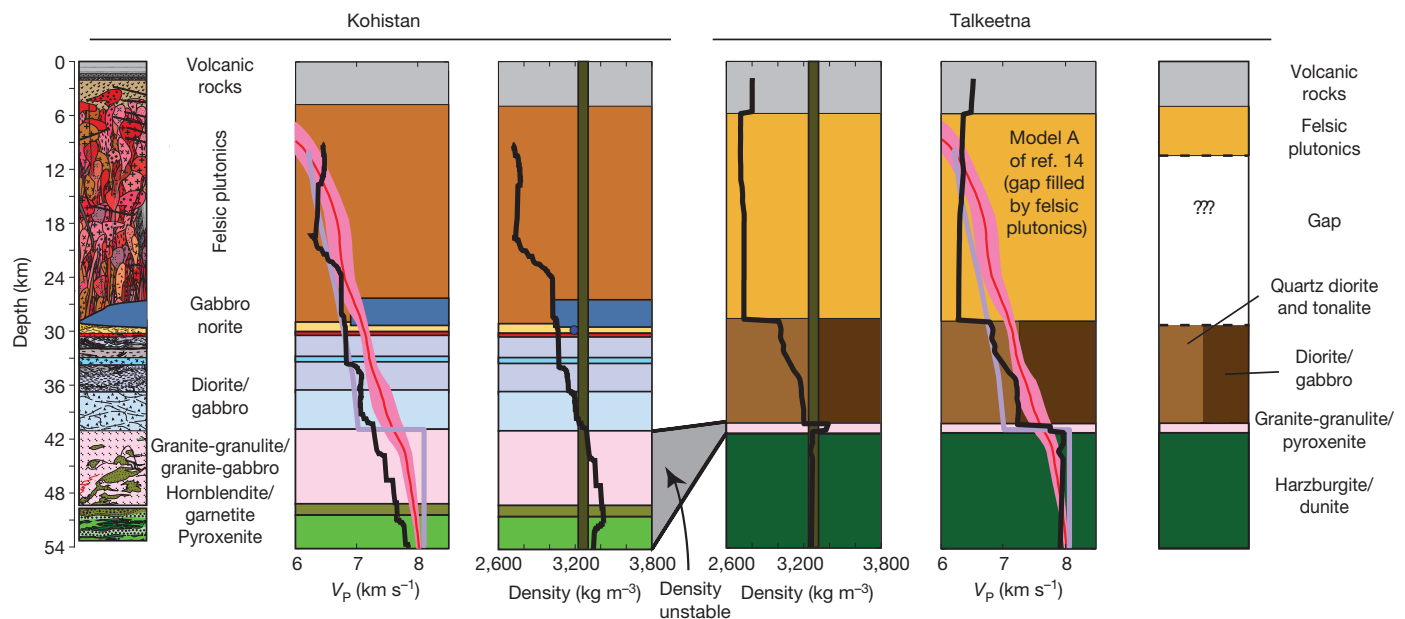


Figure 3 | Schematic illustrations of the lithological, seismic and density properties of the Kohistan and Talkeetna arc sections. Shown are simplified, schematic crustal columns (after refs 19, 22 and 33) and the calculated average seismic V_p velocities and densities (black lines) of the main crustal building blocks of the two arcs. The thickness of the different units was approximated

using the calculated densities and existing barometric pressure estimates^{19,22,33}. The pink and purple lines indicate the seismic velocities of the Izu-Bonin and continental arc crust, respectively (as in Fig. 2). The V_p and V_s estimates of Kohistan are after Fig. 2, and those of Talkeetna are recalculated after ref. 14.

for the transformation of an arc-type Moho to a continental-type Moho.

Density sorting of the lower crust can occur during collision, tectonic underplating and/or normal arc buildup²³. To constrain the maximum thickness that an unstable layer can achieve before foundering occurs, we calculated the timescale for the initiation of a Rayleigh–Taylor-type instability at the base of the arc crust as a function of the density and temperature of the underlying mantle and the density and thickness of

the unstable layer^{13,24–26} and compared it to the timescale for crustal growth for different magma supply rates¹⁰ (see Methods for details). For a given temperature, we assume that the unstable layer will grow until the timescale for instability initiation is less than the time required to form the layer. For temperatures below about 700–800 °C, instability times exceed reasonable geological timescales, because the high viscosity of the ‘cold’ crust and underlying mantle inhibits instability growth (Fig. 4). In contrast, for temperatures over 800 °C, instability growth becomes more efficient and the unstable layer grows to a thickness of only a few kilometres before foundering into the underlying mantle (on a timescale of 0.5–5 million years). The difference in the observed thermal regime between Kohistan and Talkeetna (Fig. 1c) is consistent with the preservation of a thick layer of density-unstable material at the base of the Kohistan arc crust, whereas in the warmer Talkeetna arc foundering is predicted to be more efficient, resulting in the preservation of a significantly thinner unstable layer (Fig. 4).

Our results show that foundering can explain both the location and primary seismic characteristics of the continental Moho. In active arcs, an unstable layer removed by foundering will be rebuilt within a few million years and so the chance of seismically imaging a newly density-sorted lower crust with a sharp Moho at about 40 km is low. After magmatism ceases at an arc, Moho temperatures will remain high until the geotherms conductively relax¹⁶. As long as the Moho temperature remains above about 700 °C, foundering will continue, but the foundering layer will not be rebuilt. Instead, it will be replaced by upper-mantle rocks, resulting in the formation of a density-sorted continental lower-crust/upper-mantle interface with a sharp Moho discontinuity. We speculate that unusually thick Archaean continental crust with a preserved 5–10-km-thick transitional zone between crust and mantle represents lower crust that has not been density sorted²⁷. More detailed seismic studies of stable continental regions are needed to test the abundance of such preserved relicts.

METHODS SUMMARY

Calculation of density and seismic velocity. We used *Perple_X* (ref. 28) to calculate subsolidus thermodynamic phase equilibria for a range of whole-rock compositions from the Kohistan¹¹ and Talkeetna^{22,29} arcs assuming 1 wt% H_2O . Seismic velocities and densities of the stable mineral assemblage and mode were calculated

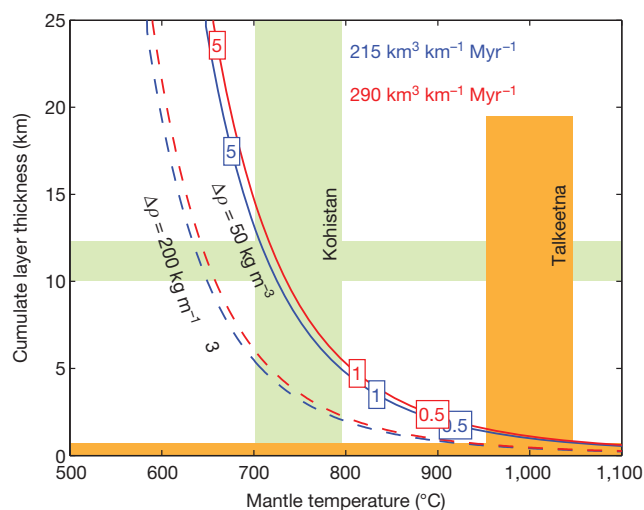


Figure 4 | Modelled thickness of the density-unstable layer at the base of arc crust. The thickness was calculated by equating the timescale required for an instability to form¹⁴ with the timescale required to grow a cumulate layer based on estimated magma fluxes¹⁰ (blue and red curves, respectively). The boxed numbers are times required to grow the layer in millions of years. Layer growth assumes that 70% of the original melt mass is partitioned into the cumulate layer¹⁰. Solid and dashed curves are based on different density contrasts between the layer and underlying mantle. The vertical bands indicate the approximate Moho temperatures for the Talkeetna (orange) and Kohistan (green) arcs, and horizontal fields indicate the preserved thickness of the density-unstable layer in the two arcs.

using a compilation of geophysical mineral properties³⁰. We implemented an updated version of the compilation (B. Hacker and G. Abers, personal communication, 2010) into *Perple_X*. The variable intrusion pressures of the rocks studied are from refs 11, 19 and 31 for Kohistan and Talcetna, respectively; corresponding depths were calculated by integrating the calculated density profiles for pressure. Temperatures were calculated along a 60 mW m⁻² geotherm for the Kohistan arc³¹ and a 80 mW m⁻² geotherm for Talcetna¹⁹. We used the following solid solution models: Atg (HP), Ctd (HP), Cpx (HP), Ep (HP), GlTrPg, Gt (HP), Pheng (HP), O (HP), Opx (HP), Pl (h), San, Sp, and T. To investigate the influence of variable oxygen fugacity (f_{O_2}) on the seismic velocity structure of an arc, we calculated seismic properties and densities at Fe^{3+}/Fe_{total} values of 0.15, 0.25 and 0.35. All results were plotted with $Fe^{3+}/Fe_{tot} = 0.25$ but the results discussed here are not dependent on f_{O_2} .

Calculation of instability timescales. Our calculation of the thickness of the unstable layer follows the approach of ref. 13. The timescale required to form an instability scales inversely with the thickness of the dense layer and the density contrast between the layer and underlying mantle (that is, thicker layers and greater density contrasts lead to shorter instability times)^{24,25}. For temperature-dependent viscosity the instability time decreases exponentially with increasing mantle temperature. For a given temperature, we assume that the unstable layer will grow until the timescale for instability initiation is less than the time required to form the layer.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 April; accepted 3 October 2013.

- Taylor, S. R. & McLennan, S. M. *The Continental Crust: Its Composition and Evolution* (Blackwell, 1985).
- Holbrook, W. S., Lizarralde, D., McGeary, S., Bangs, N. & Diebold, J. Structure and composition of the Aleutian island arc and implications for continental crustal growth. *Geology* **27**, 31–34 (1999).
- Calvert, A. in *Arc–Continent Collision* 87–119 (Springer, 2011).
- Christensen, N. I. & Mooney, W. D. Seismic velocity structure and composition of the continental crust: a global view. *J. Geophys. Res.* **B 100**, 9761–9788 (1995).
- Takahashi, N., Kodaira, S., Tatsumi, Y., Kaneda, Y. & Suyehiro, K. Structure and growth of the Izu-Bonin-Mariana arc crust: 1. Seismic constraint on crust and mantle structure of the Mariana arc-back-arc system. *J. Geophys. Res.* **113**, B01104 (2008).
- Tatsumi, Y. *et al.* Structure and growth of the Izu-Bonin-Mariana arc crust: 2. Role of crust-mantle transformation and the transparent Moho in arc crust evolution. *J. Geophys. Res.* **113**, B02203 (2008).
- Shillington, D. J., Van Avendonk, H. J. A., Holbrook, W. S., Kelemen, P. B. & Hornbach, M. J. Composition and structure of the central Aleutian island arc from arc-parallel wide-angle seismic data. *Geochem. Geophys. Geosyst.* **5**, Q10006 (2004).
- Kay, R. W. & Mahburg Kay, S. Delamination and delamination magmatism. *Tectonophysics* **219**, 177–189 (1993).
- Arndt, N. T. & Goldstein, S. L. An open boundary between lower continental crust and mantle: its role in crust formation and crustal recycling. *Tectonophysics* **161**, 201–212 (1989).
- Jagoutz, O. & Schmidt, M. W. The nature and composition of the crustal delamination in arcs. *Earth Planet. Sci. Lett.* **371–372**, 177–190 (2013).
- Jagoutz, O. & Schmidt, M. W. The formation and bulk composition of modern juvenile continental crust: the Kohistan arc. *Chem. Geol.* **298–299**, 79–96 (2012).
- Sobolev, S. V. & Babeyko, A. Y. Phase transformations in the lower continental crust and its seismic structure. *Geophys. Monogr. Ser.* **51**, 311–320 (1989).
- Behn, M. D., Hirth, G. & Kelemen, P. B. Trench-parallel anisotropy produced by foundering of arc lower crust. *Science* **317**, 108–111 (2007).
- Behn, M. D. & Kelemen, P. B. Stability of arc lower crust: insights from the Talcetna arc section, south central Alaska, and the seismic structure of modern arcs. *J. Geophys. Res.* **111**, B11207 (2006).
- Jagoutz, O., Muentener, O., Schmidt, M. W. & Burg, J. P. The respective roles of flux- and decompression melting and their relevant liquid lines of descent for Continental Crust formation: evidence from the Kohistan arc. *Earth Planet. Sci. Lett.* **303**, 25–36 (2011).
- Kelemen, P., Rilling, J. L., Parmentier, E. M., Mehl, L. & Hacker, B. R. in *Inside Subduction Factory* Vol. 138, 293–311 (ed. Eiler, J. M.) (Geophysical Monograph, American Geophysical Union, 2003).
- Shillington, D. J., Van Avendonk, H. J. A., Behn, M. D., Kelemen, P. B. & Jagoutz, O. Constraints on the composition of the Aleutian arc lower crust from V_P/V_S . *Geophys. Res. Lett.* **40**, 2579–2584 (2013).
- Ringuette, L., Martignole, J. & Windley, B. F. Magmatic crystallization, isobaric cooling, and decompression of the garnet-bearing assemblages of the Jijal Sequence (Kohistan Terrane, western Himalayas). *Geology* **27**, 139–142 (1999).
- Hacker, B. R. *et al.* Reconstruction of the Talcetna intraoceanic arc of Alaska through thermobarometry. *J. Geophys. Res.* **113**, B03204 (2008).
- Miller, D. J. & Christensen, N. I. Seismic signature and geochemistry of an island arc: a multidisciplinary study of the Kohistan accreted terrane, northern Pakistan. *J. Geophys. Res.* **B 99**, 11623–11642 (1994).
- Kono, Y., Ishikawa, M., Harigane, Y., Michibayashi, K. & Arima, M. P- and S-wave velocities of the lowermost crustal rocks from the Kohistan arc: implications for seismic Moho discontinuity attributed to abundant garnet. *Tectonophysics* **467**, 44–54 (2009).
- Kelemen, P., Hanghoj, K. & Greene, A. in *The Crust* Vol. 3 *Treatise on Geochemistry* (ed. Rudnick, R. L.) 593–659 (Elsevier-Pergamon, 2003).
- Hacker, B. R., Kelemen, P. B. & Behn, M. D. Differentiation of the continental crust by reamination. *Earth Planet. Sci. Lett.* **307**, 501–516 (2011).
- Jull, M. & Kelemen, P. B. On the conditions for lower crustal convective instability. *J. Geophys. Res.* **B 106**, 6423–6446 (2001).
- Conrad, C. P. & Molnar, P. The growth of Rayleigh-Taylor-type instabilities in the lithosphere for various rheological and density structures. *Geophys. J. Int.* **129**, 95–112 (1997).
- Conrad, C. P., Behn, M. D. & Silver, P. G. Global mantle flow and the development of seismic anisotropy: differences between the oceanic and continental upper mantle. *J. Geophys. Res.* **112**, B07317 (2007).
- Guggisberg, B., Kaminski, W. & Prodehl, C. Crustal structure of the Fennoscandian shield: a traveltimes interpretation of the long-range FENNOLOGA seismic refraction profile. *Tectonophysics* **195**, 105–137 (1991).
- Connolly, J. A. D. Computation of phase equilibria by linear programming: a tool for geodynamic modeling and its application to subduction zone decarbonation. *Earth Planet. Sci. Lett.* **236**, 524–541 (2005).
- Greene, A. R., DeBari, S. M., Kelemen, P., Blusztajn, J. S. & Clift Peter, D. A detailed geochemical study of island arc crust: the Talcetna arc section, south-central Alaska. *J. Petrol.* **47**, 1051–1093 (2006).
- Hacker, B. R. & Abers, G. A. Subduction factory 3: an Excel worksheet and macro for calculating the densities, seismic wave speeds, and H₂O contents of minerals and rocks at pressure and temperature. *Geochem. Geophys. Geosyst.* **5**, Q01005 (2004).
- Jagoutz, O. *et al.* TTG-type plutonic rocks formed in a modern arc batholith by hydrous fractionation in the lower arc crust. *Contrib. Mineral. Petrol.* **166**, 1099–1118 (2013).
- Kodaira, S. *et al.* New seismological constraints on growth of continental crust in the Izu-Bonin intra-oceanic arc. *Geology* **35**, 1031–1034 (2007).
- Jagoutz, O. & Schmidt, M. W. The formation and bulk composition of modern juvenile continental crust: the Kohistan arc. *Chem. Geol.* **298–299**, 79–96 (2012).

Acknowledgements The work was supported by NSF grant numbers EAR 0910644 (to O.J.) and EAR 1316333 (to M.D.B.). We thank N. Arndt for comments that helped to improve the manuscript. J. Connolly's help in recalibrating the elastic property calculation of *Perple_X* is appreciated, as are discussions with P. Kelemen and B. Hacker.

Author Contributions O.J. designed the project. Both authors conducted the calculations, contributed to the interpretation of the results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.J. (jagoutz@mit.edu).

METHODS

The density of arc lower crust rocks. Crustal rocks that are denser than the underlying upper mantle peridotite are considered density unstable. Density-unstable rocks can either form as relatively Fe-rich ultramafic cumulates derived from mantle-derived melts, which are as dense or slightly denser than the underlying mantle at magmatic temperature but can become significantly denser upon cooling^{9,41}. Additionally, dense garnet-bearing cumulates can form from hydrous basaltic-andesitic liquids at high pressures (above about 1 GPa)⁴². Three important pressure-dependent metamorphic reactions result in the formation of dense minerals (such as spinel and garnet), which strongly control the density of Fe-rich and Al-rich (such as gabbroic) compositions in the arc lower crust.

The three main densification reactions likely to be important in the lower arc crust are: (1) The breakdown of plagioclase next to olivine, which occurs at pressures of about 0.6–0.7 GPa (ref. 43): olivine + plagioclase → pyroxene + spinel (I) (2) The formation of metamorphic garnet due to the breakdown of plagioclase at 0.8–1 GPa (ref. 43): plagioclase + orthopyroxene → garnet + quartz (II) (3) The breakdown of plagioclase at about 1.2–1.6 GPa: albite → jadeite + quartz (III).

The importance of each reaction for densification depends on the bulk composition of the system. Reaction (I) is important for olivine and plagioclase-rich rocks (troctolite and olivine-gabbro)¹⁵, which probably form in thin arcs where magma fractionation occurs at shallower crustal levels and the olivine + plagioclase stability field is increased⁴². Reaction (II) is important for rocks with high Fe/Mg ratios and high Al-content and low Si-content, such as cumulates formed from hydrous arc magmas at increased pressures¹⁵. Reaction (III) will only be relevant for strongly over-thickened arc crust.

Additionally, the depth range corresponding to pressures of about 0.8–1 GPa at which reaction (II) occurs varies significantly depending on the density structure of the arc crust. In juvenile arcs, where most of the arc crust is composed of rocks with approximately basaltic compositions with densities of around 2,900–3,100 kg m⁻³, pressures of 0.8–1 GPa correspond to depths of about 26–34 km. In contrast, in mature arcs that have a significant thickness of granitic upper crust (such as the Izu-Bonin arc) with densities as low as 2,600–2,800 kg m⁻³, pressures of 0.8–1 GPa can correspond to depths of up to 32–40 km. Accordingly, the depth range in which delamination—owing to the formation of magmatic/metamorphic garnet and/or pyroxene and spinel—occurs is 20–70 km, depending in detail on the composition of the rocks in the arc crust.

Geological setting. The Kohistan arc, exposed in northeast Pakistan, was a long-lived Jurassic/Cretaceous to Tertiary island arc that formed in the equatorial part of the Neotethyan ocean separating India and Eurasia before the India–Asia collision. The Kohistan arc exposes a complete arc section ranging from unmetamorphosed sediments in the north to upper-mantle rocks in the south. Pressure and temperature estimates for the lowermost mafic arc crust indicate pressures in excess of about 1.5 GPa for the crust–mantle transition.

The Talkeetna arc, exposed in south central Alaska, is a Triassic island arc that was active from about 200–175 million years ago. It exposes rocks ranging from unmetamorphosed sediments and associated volcanics in the north of the arc to upper-mantle rock in the south of the arc. Owing to large-offset strike–slip faulting, the middle crust is partly missing¹⁹. The lowermost mafic arc crust records maximum pressures of 1.0–1.1 GPa, indicating a slightly shallower crust–mantle transition in the Talkeetna compared to the Kohistan.

Calculation of density and seismic velocity. We used Perple_X (ref. 28) to calculate subsolidus thermodynamic phase equilibria for a wide range of whole-rock compositions from the Kohistan¹¹ and Talkeetna^{22,29} arcs assuming 1 wt% H₂O. Seismic velocities (V_p , V_s) and densities of the stable mineral assemblage and mode were calculated using a compilation of geophysical mineral properties³⁰. We implemented an updated version of the compilation (B. Hacker and G. Abers, personal communication, 2010) into Perple_X. The variable intrusion depths of the rocks studied are from refs 11, 19, 31 and 44 for Kohistan and Talkeetna, respectively. Temperatures were constrained along a 60 mW m⁻² geotherm constrained for the Kohistan arc^{31,44} and a 80 mW m⁻² geotherm for the Talkeetna arc¹⁹. We used the following solid solution models in our calculation: Atg, Chl (HP), Ctd (HP), Cpx (HP), Ep (HP), GlTrPg, Gt (HP), Pheng (HP), O (HP), Opx (HP), Pl (h), San, Sp and T.

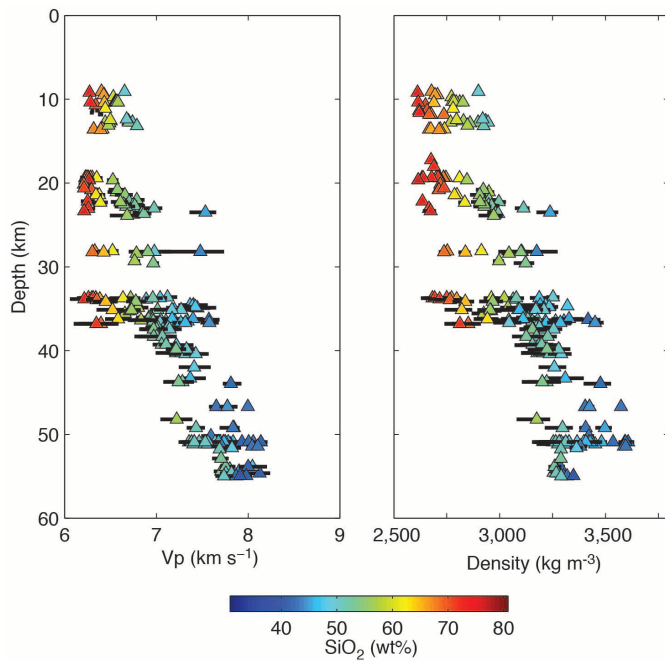
To investigate the influence of variable f_{O_2} on the seismic velocity structure of an arc, we calculated seismic properties and densities at Fe^{3+}/Fe_{total} values of 0.15, 0.25 and 0.35. All results were plotted with $Fe^{3+}/Fe_{total} = 0.25$ (ref. 45) but the results discussed here are not dependent on f_{O_2} .

The effect of temperature on the density structure. The thermal regime in the lower crust is poorly constrained and probably highly variable through time owing to the intrusion of hot basaltic liquids. To evaluate the effect of variable temperature we calculated the density and seismic structure of the Kohistan and Talkeetna crust at 40, 60 and 80 mW m⁻² geotherms (Extended Data Fig. 1). Because magmatic and metamorphic phase boundaries involving significant volume changes (and corresponding density changes) are dominantly pressure-dependent, and only to a limited extent temperature-dependent, the density structure is only marginally influenced by the thermal structure. The most important reaction at higher temperature is the breakdown of hydrous phases (for example, amphibole), which generally break down to a denser phase (for example, pyroxene). However, this transformation has only a limited effect on density and seismic properties (Extended Data Fig. 1).

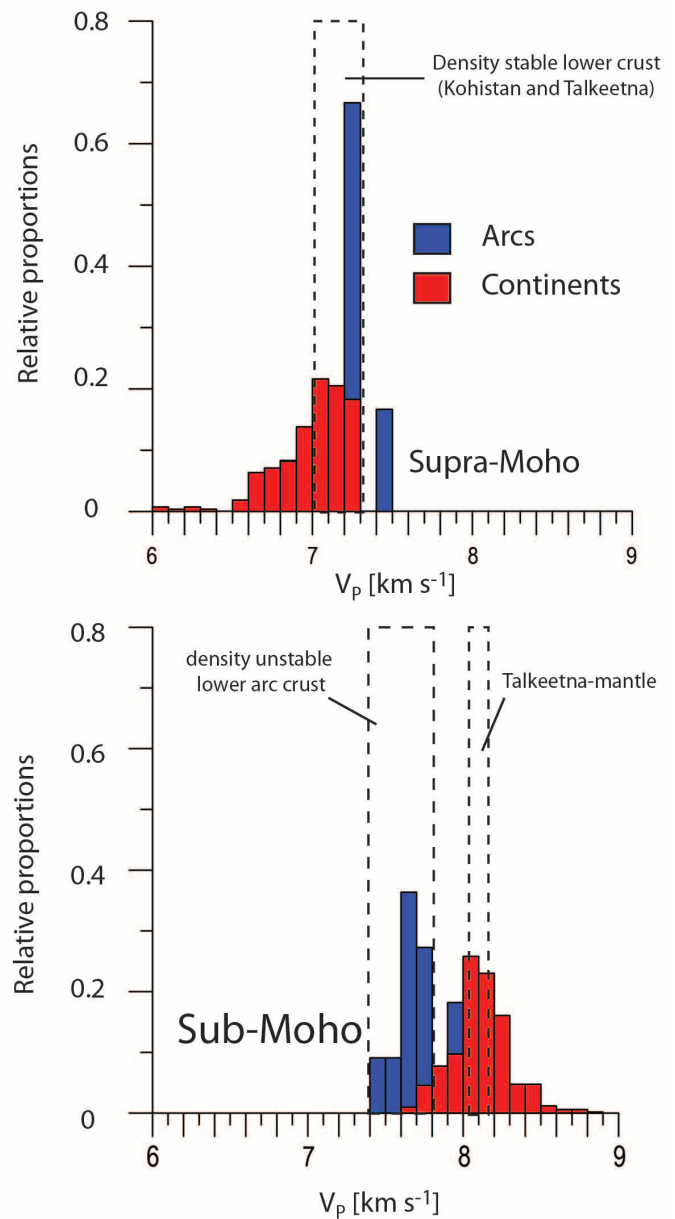
Calculation of instability timescales. Our calculation of the thickness of the unstable layer follows the approach of ref. 13. The timescale required for an instability to form scales inversely with the thickness of the dense layer, and the density contrast between the layer and underlying mantle (that is, thicker layers and greater density contrasts lead to shorter instability times)^{24,25}. In addition, for temperature-dependent viscosity the instability time decreases exponentially with increasing mantle temperature. For a given temperature, we assume that the unstable layer will grow until the time required for instability initiation is less than the time required to form the layer.

References for Fig. 1. The V_p/V_s and V_p estimates for different arcs in Fig. 1 are taken from refs 46–49.

34. Nakanishi, A. *et al.* Crustal evolution of the southwestern Kuril Arc, Hokkaido Japan, deduced from seismic velocity and geochemical structure. *Tectonophysics* **472**, 105–123 (2009).
35. Kopp, H. *et al.* Deep structure of the central Lesser Antilles Island Arc: relevance for the formation of continental crust. *Earth Planet. Sci. Lett.* **304**, 121–134 (2011).
36. Iwasaki, T. *et al.* Crustal and upper mantle structure in the Ryukyu Island Arc deduced from deep seismic sounding. *Geophys. J. Int.* **102**, 631–651 (1990).
37. Iwasaki, T. *et al.* Precise P and S wave velocity structures in the Kitakami massif, Northern Honshu, Japan, from a seismic refraction experiment. *J. Geophys. Res.* **99**, 22187–22204 (1994).
38. Kodaira, S. *et al.* Seismological evidence for variable growth of crust along the Izu intraoceanic arc. *J. Geophys. Res.* **112**, B05104 (2007).
39. Takahashi, N. *et al.* Crustal structure and evolution of the Mariana intra-oceanic island arc. *Geology* **35**, 203–206 (2007).
40. Calvert, A. J., Klemperer, S. L., Takahashi, N. & Kerr, B. C. Three-dimensional crustal structure of the Mariana island arc from seismic tomography. *J. Geophys. Res.* **113**, B01406 (2008).
41. Müntener, O. & Ulmer, P. Experimentally derived high-pressure cumulates from hydrous arc magmas and consequences for the seismic velocity structure of lower arc crust. *Geophys. Res. Lett.* **33**, L21308 (2006).
42. Müntener, O., Kelemen, P. B. & Grove, T. L. The role of H₂O during crystallization of primitive arc magmas under uppermost mantle conditions and genesis of igneous pyroxenites; an experimental study. *Contrib. Mineral. Petrol.* **141**, 643–658 (2001).
43. Kushiro, I. & Yoder, H. S. Jr. Anorthite-forsterite and anorthite-enstatite reactions and their bearing on the basal-eclogite transformation. *J. Petrol.* **7**, 337–362 (1966).
44. Jagoutz, O. The fine scale seismic structure of an exposed island arc section based on field and petrological constraints. *AGU Fall Meet. Abstr.* 2628 (2011).
45. Cottrell, E. & Kelley, K. A. The oxidation state of Fe in MORB glasses and the oxygen fugacity of the upper mantle. *Earth Planet. Sci. Lett.* **305**, 270–282 (2011).
46. Zhang, H. *et al.* High-resolution subducting-slab structure beneath northern Honshu, Japan, revealed by double-difference tomography. *Geology* **32**, 361–364 (2004).
47. Syracuse, E. M. *et al.* Seismic tomography and earthquake locations in the Nicaraguan and Costa Rican upper mantle. *Geochem. Geophys. Geosyst.* **9**, <http://dx.doi.org/10.1029/2008GC001963> (2008).
48. Eberhart-Phillips, D. *et al.* Imaging the transition from Aleutian subduction to Yakutat collision in central Alaska, with local earthquakes and active source data. *J. Geophys. Res.* **111**, <http://dx.doi.org/10.1029/2005JB004240> (2006).
49. Wang, Z. & Zhao, D. V_p and V_s tomography of Kyushu, Japan: New insight into arc magmatism and forearc seismotectonics. *Phys. Earth Planet. Inter.* **157**, 269–285 (2006).



Extended Data Figure 1 | Seismic velocity and density along different geotherms for the Kohistan arc. Plotted are the mean and range in V_p and density as calculated along the 40, 60 and 80 mW m^{-2} geotherms.



Extended Data Figure 2 | Seismic velocities of the lower arc and continental crust. Histogram showing distribution of average seismic velocities directly above and below the Moho in continents (red, after ref. 4) and from active arcs (refs 6, 8, 32, 34–40). Also shown are the range of V_p for density-stable and density-unstable rocks from the Kohistan and Talkeetna arcs, as dashed fields calculated from this study. In the arcs, sub-Moho rocks have on average a V_p that is 0.5 km s^{-1} slower than do sub-Moho rocks in continents. The observed low velocities in the arcs agree with the velocities calculated for density-unstable crustal rocks from Kohistan.

Genetic incompatibilities are widespread within species

Russell B. Corbett-Detig¹, Jun Zhou¹, Andrew G. Clark^{2,3}, Daniel L. Hartl¹ & Julien F. Ayroles^{1,2,4}

The importance of epistasis—non-additive interactions between alleles—in shaping population fitness has long been a controversial topic, hampered in part by lack of empirical evidence^{1–4}. Traditionally, epistasis is inferred on the basis of non-independence of genotypic values between loci for a given trait. However, epistasis for fitness should also have a genomic footprint^{5–7}. To capture this signal, we have developed a simple approach that relies on detecting genotype ratio distortion as a sign of epistasis, and we apply this method to a large panel of *Drosophila melanogaster* recombinant inbred lines^{8,9}. Here we confirm experimentally that instances of genotype ratio distortion represent loci with epistatic fitness effects; we conservatively estimate that any two haploid genomes in this study are expected to harbour 1.15 pairs of epistatically interacting alleles. This observation has important implications for speciation genetics, as it indicates that the raw material to drive reproductive isolation is segregating contemporaneously within species and does not necessarily require, as proposed by the Dobzhansky–Muller model, the emergence of incompatible mutations independently derived and fixed in allopatry. The relevance of our result extends beyond speciation, as it demonstrates that epistasis is widespread but that it may often go undetected owing to lack of statistical power or lack of genome-wide scope of the experiments.

The role of epistasis in shaping genetic variation and contributing to observable differences within and between populations has been the focus of much debate^{1–3}. In complex trait genetics, the additive paradigm used in genome-wide association studies¹⁰ has recently been challenged by mounting evidence highlighting the importance of non-additive interactions between alleles⁴. Although the debate has been centred on the relative contribution of epistasis to the genetic variance, we still have a poor grasp of the extent to which epistasis affects the mean genotypic values of traits, an important step towards understanding the genetic basis of complex traits and the organization of molecular pathways⁵. Although epistasis is widely accepted to underlie the genetic basis of speciation, many details of this phenomenon remain poorly understood^{2,3,5}. In particular, the evolutionary origins of the alleles that cause reproductive isolation are largely unidentified. Therefore, the importance of epistasis in shaping fitness within and between populations remains an important question in evolutionary biology.

Our understanding of the contribution of epistasis and the molecular details underlying non-additive genetic interactions is limited largely by the scarcity of available data. Although the idea that populations may harbour alleles with epistatic fitness effects has existed in the literature for some time, very few examples have been dissected at the genetic level (except for individual cases^{6,11}). Furthermore, as yet, no systematic surveys have been conducted in diploid out-crossing species that have adequate statistical power to detect small fitness effects or to finely map interacting loci.

The traditional approach used to detect epistasis by statistical means relies on the observation of non-additivity of genotypic values between loci for a given phenotype. However, epistasis for fitness should have

a genomic signature, regardless of our ability to measure a given phenotype^{5–7}. In particular, it is expected that unfavourable allelic combinations will be under-represented, and this should precipitate a deviation from Mendelian proportions among unlinked incompatible alleles (detected by performing a screen for statistical association between alleles at loci that are not physically linked; Methods). Hereafter we refer to such deviations as genotype ratio distortion (GRD). In natural populations an exhaustive search for GRD is computationally intractable, statistically underpowered, or both⁶. By contrast, model organisms allow us to create experimental populations in which the amount of genetic variation and recombination can be controlled, thereby amplifying the signature of epistasis in a background of reduced dimensionality.

Here we apply tests of epistasis to the *Drosophila* Synthetic Population Resource (DSPR)^{8,9} (Extended Data Fig. 1). To create the DSPR, two sets of eight highly inbred strains of diverse geographic origins were independently crossed in a round-robin design. Each set was duplicated and maintained for 50 generations in large freely-mating population cages. Subsequently, approximately 400 recombinant inbred lines (RILs) in each of four independent panels were created through 20 generations of sib-mating (generating four ‘panels’: A-1, A-2 and B-1, B-2). After inbreeding, each RIL was genotyped at densely spaced markers, allowing a description of the genome of each RIL as a genetic mosaic of the eight founding lines originally crossed (Extended Data Fig. 1). The 50 generations of recombination and the large number of RILs within a panel provides replication over random allelic permutations. This replication is essential to attain statistical power for the detection of small effect epistasis.

We first excluded the possibility that residual population structure within the DSPR created association among alleles in the absence of epistasis by performing principal component analysis (Extended Data Fig. 2, Methods). Subsequently, we identified 22 pairs of epistatically interacting alleles in the DSPR (Fig. 1, Extended Data Table 1, Extended Data Fig. 3). Importantly, of the 44 incompatible alleles, 27 appear to be shared between two or more strains (Extended Data Table 1). This indicates that incompatible alleles are segregating at polymorphic frequencies in natural populations, and are not a result of inbreeding or long-term maintenance at small population size. On the basis of the frequencies in the founder strains, we estimate that any pairwise combination of founders has, on average, 1.15 pairs of epistatically interacting alleles. This is probably an underestimate, both because our statistical approach is conservative and because selectively disfavoured allelic combinations may be purged by selection during the free-recombination phase of the DSPR.

We next sought to confirm the predicted effect on reproductive fitness and to identify the underlying phenotype of two pairs of incompatible haplotypes (Fig. 2a, c). Using the original founder strains that contributed the putatively interacting alleles, we performed experimental crosses, and in both cases, we discovered that the negative interaction is caused by the minor alleles at each locus (genotype *aabb* in Fig. 2b, d; Extended Data Fig. 1). Specifically, in the case of one incompatibility

¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ²Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA. ³Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. ⁴Harvard Society of Fellows, Harvard University, Cambridge, Massachusetts 02138, USA.

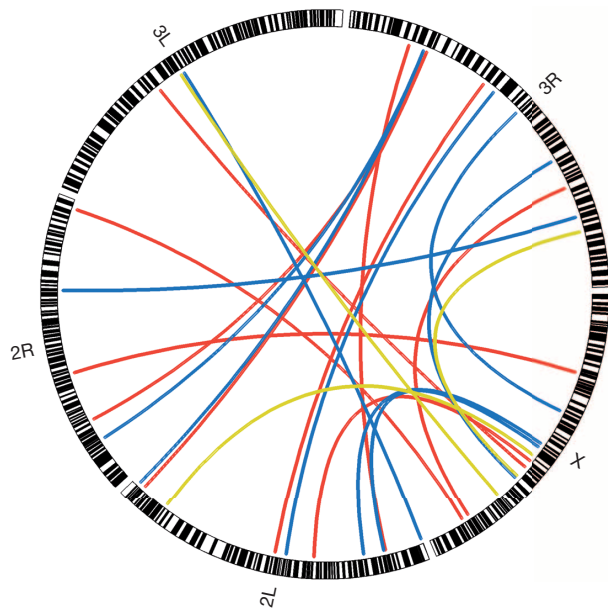


Figure 1 | Locus pairs showing significant GRD across the DSPR lines of *Drosophila*. The outer circle represent each chromosome arm. Each link represents a locus pair showing significant two-locus GRD. Yellow, blue and red links correspond respectively to RIL panel A-2, B-1 and B-2 (5% FDR corrected $P < 0.05$).

between chromosomes 2 and 3, males that are homozygous for both incompatible alleles produce on average 74% fewer offspring compared to all other allelic combinations ($P = 5.51121 \times 10^{-9}$ Likelihood Ratio Test, LRT, Fig. 2b, Extended Data Fig. 4). No significant effect was detected in females for any combination of genotypes. Using the same approach we validated a second instance of GRD, selected in the low range of effect size, between a haplotype on chromosomes X and 3 (Extended Data Fig. 5). We again observe a significant decrease (22%) in F_2 male fertility ($P = 8.25 \times 10^{-5}$ LRT, Fig. 2b, Extended Data Fig. 4), suggesting that GRD is a reliable signature of epistasis. The ‘faster-males’ theory^{2,12} and subsequent experimental confirmations (reviewed in ref. 13) predict that male infertility will evolve more rapidly than other forms of post-zygotic reproductive isolation. Although we only have phenotypic data for our confirmed examples, the fact that both implicate male fertility as the underlying phenotype suggests that this effect may extend to within-species fitness epistasis.

The DSPR was intercrossed for sufficiently many generations (in excess of 50) that little linkage disequilibrium remains; hence this approach allows us to narrow down likely candidate genes associated with epistatic interaction for male fecundity. In total, there are three genes within the haplotype on chromosome arm 2R (~40 kb). The gene *notopleural* (*np*) is at the peak of this region; it is expressed in mature sperm¹⁴ with alleles that are known to affect viability and sterility¹⁵. Notably, the human orthologue of *np* is associated with sperm-dysfunction in humans¹⁶. The interacting haplotype on chromosome arm 3R contains only two genes. In the centre of this region is *Cyp12e1*, a P450-cytochrome associated with electron transport in the mitochondria¹⁷. Interestingly, *Cyp12e1* harbours a non-synonymous mutation in a highly conserved protein domain. Mitochondrial dysfunction is commonly associated with male sterility in humans, plants and *D. melanogaster*¹⁸, and therefore seems a plausible candidate phenotype.

To confirm that these observations were not specific to the *Drosophila* DSPR, we used the same method to screen for GRD in two additional RIL panels: the MAGIC panel in *Arabidopsis*¹⁹ and the NAM panel in maize²⁰. We found 7 instances of GRDs in *Arabidopsis* and 5 in maize (Extended Data Table 2). Although we have not validated these results, they suggest that GRD is present in other species as well.

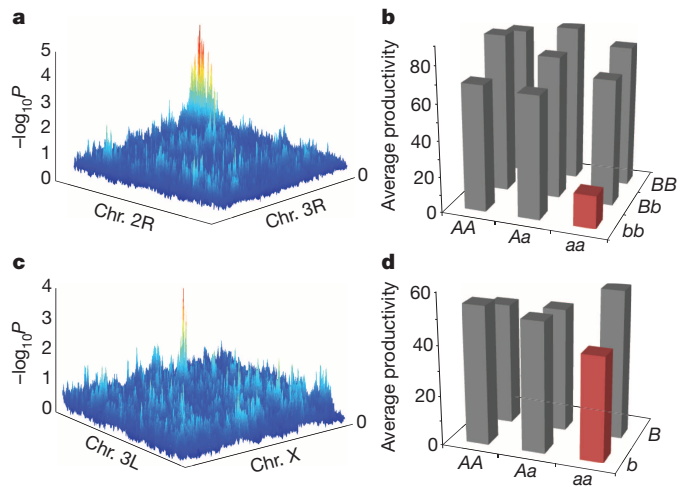


Figure 2 | From missing genotypes to epistasis. **a**, GRD signature between all genotyped loci on chromosomes (Chr.) 2R and 3R in RIL panel B-2. **b**, P value. **b**, Average productivity of each genotypic class recovered from 318 F_2 single-pair matings (progeny counts are F_3). As predicted from the GRD signal (in **a**), haplotypes tagged by single nucleotide polymorphism (SNPs) at positions 2R:4806926 and 3R:5870973 show strong negative epistasis for the *aa;bb* genotypes, $P = 5.51121 \times 10^{-9}$ LRT²⁹ (indicated by the red bar). **c**, GRD between loci on chromosomes 3L and X in RIL panel A-2. **d**, Average productivity of each genotypic class recovered from 401 F_2 single-pair matings. Haplotypes tagged by SNPs at positions 3L:11510853 and X:16483812 show strong negative epistasis for the minor alleles on each haplotype *aa;bb*, $P = 8.25 \times 10^{-5}$ LRT²⁹ (indicated by the red bar).

Although the contribution of epistasis to variation in fitness is controversial in some fields²¹, the Dobzhansky–Muller incompatibility (DMI) model^{2,22,23} is a widely accepted guiding principle for biologists studying of the genetic basis of intrinsic, post-zygotic reproductive isolation. Largely motivated by this model, which predicts that alleles causing hybrid incompatibility are derived and fixed after population divergence, much empirical work in speciation genetics has been dedicated to mapping DMIs between species that diverged relatively long ago on an evolutionarily timescale^{1,2} (Extended Data Fig. 5). However, it is unclear if these known examples of so-called ‘speciation genes’^{1,2,23,24} are an accurate representation of the earliest events in speciation, which have the greatest biological significance². Even species that have diverged for only ~250,000 years have evolved complete male sterility an estimated 15 times over²⁵. A reasonable interpretation of this evidence may concede that known ‘speciation genes’ are unlikely to be the same as those that initially contributed to reproductive isolation, but that these examples are instructive as to the properties of those genes²—an argument that closely mirrors our own.

Our central finding, that fitness epistasis is widespread within natural populations, indicates that the raw material to drive reproductive isolation is segregating contemporaneously within species and does not necessarily require, as proposed by the DMI model²², the emergence of genetically incompatible mutations independently derived and fixed in allopatric lineages²³. It is therefore necessary to explore the possibility that reproductive isolation could be achieved through divergence in frequencies of numerous pre-existing, polymorphic, small-effect incompatibilities^{26–28} (Fig. 3). The implications of the present results go beyond understanding the role of intra-specific incompatibility in the context of speciation. Our work shows that epistasis for fitness-related traits has a detectable genomic footprint, and supports the idea that latent incompatibilities often exist between segregating variation within populations, only to be released when divergent lineages hybridize. This discovery highlights the importance of understanding the contribution of epistasis to observable phenotypic differences within and between populations.

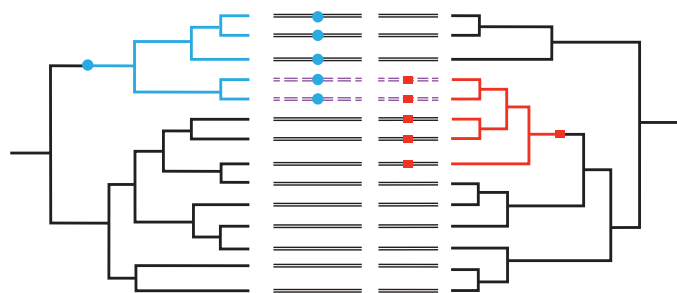


Figure 3 | Model for unlinked loci with segregating pairs of incompatible alleles. The dendrograms on the left and the right represent the genealogies of two haplotypes segregating within a species. The blue dot and the red rectangle indicate the origins of incompatible mutations on each respective genealogy. On the left, derived blue alleles are incompatible with derived red alleles on the right. These genealogies yield the individuals shown in the centre, wherein each line segment corresponds to a chromosome and each coloured square indicates the derived incompatible allele. Importantly, these incompatible allele pairs are polymorphic in this sample of individuals, thus individuals who inherit both incompatible alleles have lower fitness than those with either none or only a single incompatibility.

METHODS SUMMARY

We genotyped the RILs of the DSPR by requiring that each putative variant be supported by a minimum of five reads. All sites wherein two or more alleles are supported by five reads were discarded. We confirmed that the RIL panels were free of cryptic population structure by performing principal component analysis (Extended Data Fig. 2). We next excluded sites wherein fewer than 150 individuals have a supported genotype, where the minor allele was present in fewer than 10 individuals, or where more than 15% of individuals with data had heterozygous genotypes. Following this, we assessed statistical significance for non-independence between pairwise combinations of alleles using a χ^2 test, and applied a 5% false discovery rate (FDR) to correct for multiple testing. To reduce type 1 error, we restricted our search to inter-chromosomal comparisons and required that each putative instance of GRD be consistent with signal from adjacent variants (see Methods).

To confirm the predictions of the GRD scan, we first crossed the two DSPR founder strains that contributed the predicted interacting alleles. We then intercrossed the F_1 progeny to produce F_2 offspring. Virgin F_2 females were then individually and randomly mated to a single F_2 male. After mating for 4 days, the F_2 pairs were individually genotyped at known variable sites near the interacting alleles. We recorded the number of progeny of each pair to assay productivity. We used TaqMan kits to perform qPCR on the F_2 parents, and performed numerous statistical analyses⁵ to quantify epistatic effects as a product of genotypes at the two sites (see Supplementary Methods).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 April; accepted 17 September 2013.

Published online 6 November 2013.

1. Presgraves, D. C. The molecular evolutionary basis of species formation. *Nature Rev. Genet.* **11**, 175–180 (2010).
2. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, 2004).
3. Cutter, A. D. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends Ecol. Evol.* **27**, 209–218 (2012).
4. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nature Rev. Genet.* **5**, 618–625 (2004).
5. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* **9**, 855–867 (2008).
6. Bomblies, K. *et al.* Autoimmune response as a mechanism for a Dobzhansky–Muller-type incompatibility syndrome in plants. *PLoS Biol.* **5**, e236 (2007).

7. Payseur, B. A. & Hoekstra, H. E. Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of house mice. *Genetics* **171**, 1905–1916 (2005).
8. King, E. G., Macdonald, S. J. & Long, A. D. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics* **191**, 935–949 (2012).
9. King, E. G. *et al.* Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome Res.* **22**, 1558–1566 (2012).
10. Zuk, O. *et al.* The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
11. Bikard, D. *et al.* Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623–626 (2009).
12. Palopoli, M. F. & Wu, C. I. Genetics of hybrid male sterility between *Drosophila* sibling species: a complex web of epistasis is revealed in interspecific studies. *Genetics* **138**, 329–341 (1994).
13. Wu, C. I., Johnson, N. A. & Palopoli, M. F. Haldane's rule and its legacy: why are there so many sterile males? *Trends Ecol. Evol.* **11**, 281–284 (1996).
14. Wasbrough, E. R. *et al.* The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *J. Proteomics* **73**, 2171–2185 (2010).
15. Dockendorff, T. C., Robertson, S. E., Faulkner, D. L. & Jongens, T. A. Genetic characterization of the 44D–45B region of the *Drosophila melanogaster* genome based on an F2 lethal screen. *Mol. Gen. Genet.* **263**, 137–143 (2000).
16. Netzel-Arnett, S. *et al.* The glycosylphosphatidylinositol-anchored serine protease PRSS21 (testisin) imparts murine epididymal sperm cell maturation and fertilizing ability. *Biol. Reprod.* **81**, 921–932 (2009).
17. Kasai, S. & Tomita, T. Male specific expression of a cytochrome P450 (Cyp312a1) in *Drosophila melanogaster*. *Biochem. Biophys. Res. Commun.* **300**, 894–900 (2003).
18. Meiklejohn, C. D., Montooth, K. L. & Rand, D. M. Positive and negative selection on the mitochondrial genome. *Trends Genet.* **23**, 259–263 (2007).
19. Kover, P. X. *et al.* Multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000551 (2009).
20. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
21. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
22. Dobzhansky, T. *Genetics and the Origin of Species* (Columbia Univ. Press, 1937).
23. Orr, H. A. & Turelli, M. The evolution of postzygotic isolation: accumulating Dobzhansky–Muller incompatibilities. *Evolution* **55**, 1085–1094 (2001).
24. Presgraves, D. C. & Stephan, W. Pervasive adaptive evolution among interactors of the *Drosophila* hybrid inviability gene, *Nup96*. *Mol. Biol. Evol.* **24**, 306–314 (2007).
25. Tao, Y. *et al.* Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. *Genetics* **164**, 1383–1397 (2003).
26. Fitzpatrick, B. M. Hybrid dysfunction: population genetic and quantitative genetic perspectives. *Am. Nat.* **171**, 491–498 (2008).
27. Demuth, J. P. & Wade, M. J. On the theoretical and empirical framework for studying genetic interactions within and among species. *Am. Nat.* **165**, 524–536 (2005).
28. Reed, L. K. & Markow, T. A. Early events in speciation: polymorphism for hybrid male sterility in *Drosophila*. *Proc. Natl Acad. Sci. USA* **101**, 9009–9012 (2004).
29. Cheverud, J. M. & Routman, E. J. Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455–1461 (1995).

Acknowledgements We are grateful to T. Long, S. Macdonald and E. King for creating the DSPR and sharing the RILs and founder strains with us. We thank C. Jones, B. de Bivort, T. Sackton, S. D. Kocher, J. Grenier and N.E. Soltis for comments and discussions. We thank X. Shi for technical assistance. This work was supported by grants: NIH GM065169 and GM084236 to D.L.H., HD059060 to A.G.C., and Harvard Society of Fellows Fellowship and Harvard Milton Funds to J.F.A. R.B.C.-D. was supported by a Harvard Prize Fellowship.

Author Contributions J.F.A. conceived the idea of the project, R.B.C.-D. and J.F.A. conceived and designed experiments and analyses. R.B.C.-D. and J.F.A. conducted bioinformatics and statistical analyses; R.B.C.-D., J.F.A. and J.Z. performed experiments; J.Z. carried out molecular work; A.G.C. and D.L.H. gave analytical and conceptual advice throughout the project.

Author Information All code used and generated for this study is available upon request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.F.A. (ayroles@fas.harvard.edu).

METHODS

The DSPR resource. The *Drosophila* Synthetic Population Resource (DSPR) is described in detail elsewhere^{30,31}. This panel of recombinant inbred lines (RILs) was generated by first crossing eight highly-inbred strains in a round-robin design and subsequently maintaining in a freely-mating large population cage. Initially two sets of eight lines were independently crossed (panels A and B). Following the round-robin cross, each panel was subdivided into two replicate sets (we refer to these panels as A-1, A-2 for set A and B-1 and B-2 for set B). The geographic origin of each line is described elsewhere³⁰. After 50 generations of recombination, approximately 400 lines were inbred by 20 generations of full sib mating. The genome of the resulting RILs is a mosaic of the original eight founder strains within each set. The DSPR resource is composed of a total of four sets of approximately 400 RILs.

We downloaded the genome sequences of the founding strains as well as the SNP pileups from the DSPR website (<http://wfitch.bio.uci.edu/~dspr/index.html>). Briefly, King *et al.*³⁰ sequenced each RIL at a subset of markers selected via a restriction enzyme associated digest (RAD; hence RAD-seq)³². This is followed by sequencing of sites adjacent to restriction enzyme cut sites (SgrAI) using 100-bp single-end reads. In addition, King *et al.*³⁰ sequenced each founder genome to approximately 50× depth, allowing for accurate genotyping of the founding strains. Thus for all sites assayed via RAD-seq in the RIL panels, it is possible to identify which founders may have contributed that particular variant.

Genotypic data. We began by applying various quality filters to RAD-seq based genotypes of each RIL. Because the error properties and ascertainment bias of RAD-seq methodologies are generally poorly understood, we adopted a conservative genotyping approach to analyse the SNP data. For each site, we required that a minimum of five reads support a genotype call in order to consider that site in a given RIL. If at a given site an RIL has two genotypes supported by five or more reads, the site was considered heterozygous and excluded from any further analysis. More than 95% of genotype calls were homozygous, which is consistent with the results of King *et al.*³⁰, who reported very high rates of homozygosity, and confirming a largely successful inbreeding. We excluded all sites for which the minor allele was present in fewer than 10 lines and for which the number of heterozygous genotypes across RILs within a panel exceeded 15% of the total.

Detecting GRD. We used the resulting SNP genotype calls to search between pairs of variable sites for non-independent allelic segregation³³, which we call genotype ratio distortion (GRD).

GRD was detected by computing a χ^2 test between each pair of alleles on different chromosomes. Significant GRD reflects a deviation from expected Mendelian genotypic ratio under independent segregation between loci. In order to remain conservative, and to ensure that significant allelic pairs were not physically linked, we restricted this search to inter-chromosomal pairs. We also excluded all pairwise comparisons with fewer than 150 total genotypes or for which any allele's frequency was less than 0.05. Statistical significance was assessed for all inter-chromosomal comparisons via a χ^2 test and we applied a 5% false discovery rate (FDR) correction for multiple-testing³⁴. Subsequently, we only report pairs of alleles for which at least three adjacent SNPs are in local linkage and also show significant GRD.

Assessing frequency of incompatible alleles in founder lines. For each instance of GRD, we attempted to estimate the frequency of the alleles in the founder population. Although many of the SNPs we identified in linkage disequilibrium with incompatible haplotypes are present in more than one of the founder strains, this does not necessarily indicate that both founders contained the interacting allele. In some cases one or more founder haplotypes may not be present in the RILs. To account for this problem, we confirmed that each parental haplotype was present in the RILs by confirming that SNPs near to instances of GRD and unique to each potentially interacting founder strain are present in the RILs and that SNPs unique to each founder also show strong associations with their predicted inter-chromosomal interactions when the other founders haplotypes are excluded.

We further confirmed our inferences of GRD by searching for associations between haplotypes as identified by the hidden Markov model implemented in ref. 30 for the analysis of the DSPR. Because haplotype probabilities are 'soft' (that is, the maximum genotype probability for any one haplotype is 0.995), we only considered individuals at sites that have a greater than 95% probability of a single haplotype (probably homozygous). We searched for GRD between all possible sets of parental haplotypes at each locus (that is, all 8 individuals, 8 choose 2 pairs, 8 choose 3 trios, and so on), and we excluded all sets that contained founder haplotypes that were not represented in the RIL panel. Here again we required that minor allele frequencies of any pairwise comparisons be a minimum of 0.05. In 66.67% of cases, estimates of founder allele frequency from the maximally significant SNP and maximally significant set of haplotypes matched perfectly. For the remaining 33.33% of cases, a likely explanation is that most significant SNPs were not in perfect linkage disequilibrium with the incompatible allele in the founders' genomes. In all cases we detected a significant interaction between a similar set of

lines as we predicted based on the SNP data. Results from these analyses are summarized in Extended Data Table 2. Importantly, our results indicate that many incompatible alleles are present in more than one founder, suggesting that they were segregating in natural populations before being 'captured' within isofemale lines.

Experimental validation. In order to experimentally validate two specific instance of GRD, we sought to identify the causative phenotype underlying two of the interactions that we discovered in the SNP data (in panel B-2 between 2R:4806926 and 3R:5870973 and panel A-1 between 3L:11510853 and X:16272168). The first instance was chosen for the strength of the interaction and the interesting biology associated with the gene harboured by each haplotype. The second one was chosen at random but aimed to represent the average magnitude of disequilibrium we found across all the interacting pairs uncovered. S. MacDonald provided the founder strains used in the construction of the DSPR.

For both incompatibilities, we initially crossed the two strains that contributed the interacting haplotypes. We then inter-crossed the F₁ progeny to produce F₂s. Both parental and F₁ crosses were performed using five males and five virgin females per vial. We collected 318 (2R–3R interaction) and 401 (3L–X interaction) virgin F₂ females and maintained them in female-only vials for between four and seven days. Males were kept for the same time in male-only vials. If after this time we did not observe any larvae, each female was then mated individually to a single F₂ male. After four days, we removed the parents and extracted DNA from each F₂ individually. Experimenters were blind to the fly genotypes until the end of the experiment. At six and twelve days after removing the parents, we cleared each vial and recorded the number of offspring that had been produced. All crosses were performed on standard medium supplemented with yeast. We maintained all vials for crosses and for ageing virgin flies on a 12 h light:dark cycle at a constant 25 °C.

Following mating, we ground single flies in 50 µl of appropriate buffer (10 mM Tris, pH 8.0; 1 mM EDTA; 25 mM NaCl; 0.5% SDS). 2 µl protease K (20 mg ml⁻¹) were added and the samples were incubated for 30 min at 37 °C, and 5 min at 95 °C. Genotyping was performed using Taqman genotyping assays at sites that differentiated the founder strains near to the instances of GRD (catalogue no. 4371353, Applied Biosystems). This assay uses two probes that differ at the SNP site of interest, with each probe complementary to one allele. We used a 5 µl reaction volume containing 2.25 µl DNA, 2.5 µl TaqMan master mix, and 0.25 µl Custom TaqMan probes. We placed a 384-well plate in the Applied Biosystems 7900HT fast real-time PCR system for qPCR reactions. The program began with a step at 95 °C for 10 min, following 40 cycles of 15 s at 92 °C and 1 min at 60 °C. We assigned genotypes to individuals using TaqMan Genotyper Software V1.3. We required that each genotype have a minimum 0.95 posterior probability of being correctly called before productivity analyses.

Various definitions of epistasis have led to many approaches to statistical detection of epistatic effects (reviewed in ref. 35). Given that epistasis will be present when the combined effect of a particular pair of alleles is different from what would be expected under additivity (that is, whether alleles at loci A and B were considered together or independently, the phenotypic effects would be equivalent), many tests rely on detecting departures of the means of the genotypic classes.

We tested for the presence of statistical epistasis by implementing the method of Cheverud and Routman³⁶ who proposed an intuitive approach which consists of fitting a linear model containing additive, dominance and interaction effects—tested against the null in which there are only additive effects. The model for autosomal loci is of the following form:

$$Y_{ijkl} = a_{ij} + a_{kl} + d_{ij} + d_{kl} + a_{ij}a_{kl} + a_{ij}d_{kl} + a_{kl}d_{ij} + d_{ij}d_{kl} + e$$

Here, Y_{ijkl} corresponds to the productivity value of flies with genotype ij at loci 1 and kl at loci 2; a_{ij} and a_{kl} for the additive effects of loci 1 and 2, respectively; d_{ij} and d_{kl} for the dominance effects of loci 1 and 2; $a_{ij}a_{kl}$ for the additive by additive epistatic effects between loci 1 and 2; $a_{ij}d_{kl}$ and $a_{kl}d_{ij}$ for the additive by dominant effects and finally $d_{ij}d_{kl}$ for the dominance by dominance effects (e , the residual error). We can then perform a likelihood ratio test to ask whether a model including interaction effect provides a better fit to the data. The model fitted to the incompatibility involving a hemizygous chromosome (between chromosomes X and 3L) was adjusted accordingly to reflect the absence of dominance and dominance interaction on the X.

For the incompatibility between chromosome 2R and 3R, we obtained significant evidence for epistasis ($P = 5.51121 \times 10^{-9}$; LRT) definitions above. Similarly, the incompatibility between chromosomes X and 3L was also significant ($P = 8.25 \times 10^{-5}$; LRT).

As we have shown, recombinant inbred lines are a powerful tool for detecting epistasis statistically. Nonetheless, the additional test-crosses we used to validate two instances of GRD were necessary to determine the role of dominance in each case of epistasis, something that cannot be assessed solely from inbred lines.

Assembly methods. To identify candidate polymorphisms in regions of GRD between 2R and 3R, we obtained short-read data for the founder strains from

the DSPR website (<http://wftch.bio.uci.edu/~dspr/>). We aligned all data to the *D. melanogaster* reference genome v5.42³⁷ using stampy³⁸, which first maps reads using BWA³⁹, and subsequently attempts to map those reads which BWA fails to confidently map under more permissive conditions. All alignments were performed using the default parameters of each program. We first realigned indels in each line using the 'RealignerTargetCreator' and 'IndelRealigner' tools contained within the Genome Analysis Toolkit (GATK⁴⁰). We called the consensus for each line using the GATK genotyping function, 'UnifiedGenotyper'. Genotyping was performed using the default parameters of the program except that we called each line as a haploid genome (`—sample_ploidy 1`), which is justified because all strains are highly inbred in the regions we focused on³⁰. Finally, we predicted the effect of all discovered variants in regions of strong GRD using SnpEff⁴⁰.

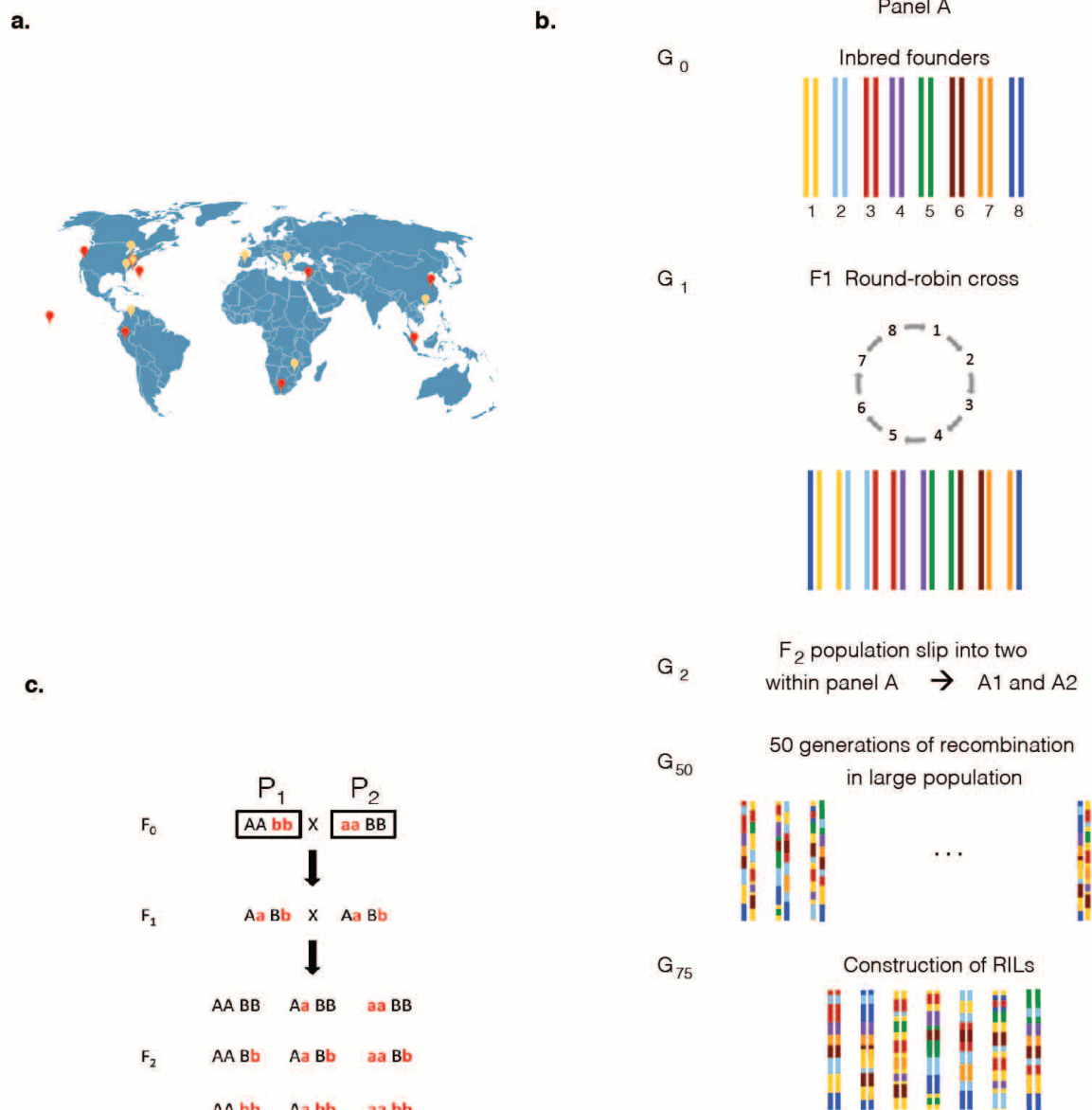
Homology search. We used blastp³⁵ to query the sequence of Cyp12e1 against the non-redundant protein database maintained by the NCBI. We used COBALT⁴¹ to align these protein sequences and to identify highly conserved protein domains. To determine if the particular mutation that we observed is unusual at this site, we counted the number of occurrences of this residue among the top 250 most closely related proteins.

Maize incompatibilities. We searched for GRD in *Zea mays* using RIL genotype data generated for the Nested-Association Mapping panel⁴². This panel is composed of 25 sets of RILs made up of 200 RILs each. Each set was produced by crossing a single strain against one of 25 other highly-inbred parents, followed by several generations of self-fertilization. Known single nucleotide variants and short repeats were then assayed using by PCR to establish the haplotype structure of each RIL.

For these analyses, we used the 'imputed' genotype calls (described elsewhere⁴²), as the 'raw' genotype data may be unreliable (E. Buckler, personal communication). We acquired the imputed data for the RILs from www.panzea.org. We then filtered data following a similar procedure to that we implemented for the *Drosophila* data (described above). We required that each site in any one individual be homozygous for consideration in downstream analyses. We further required each allele in any pairwise comparison have a minimum frequency of 0.05. Statistical significance was assessed using a χ^2 test and corrected for multiple testing by applying a 10% FDR correction within each set of RILs³⁴. Here we used a slightly higher FDR than for the *Drosophila* data because the NAM RIL panels are smaller and therefore have reduced power to detect interacting alleles.

***Arabidopsis* incompatibilities.** The multi-parental advanced generation intercross panel of *Arabidopsis thaliana* RILs⁴³ was produced in a similar way to the DSPR panel. Initially 18 founder strains were intercrossed followed by several generations of self-fertilization to produce approximately 500 mostly independent RILs. Each RIL was then genotyped at known variable sites using PCR⁴³. Once again, in analysing these data, we required that each site be homozygous in order to be considered in downstream analyses and that each allele within any pairwise comparison be present at a frequency greater than 0.05. Afterwards, we identified significant GRD between haplotypes using χ^2 tests and applied a 10% FDR correction³³.

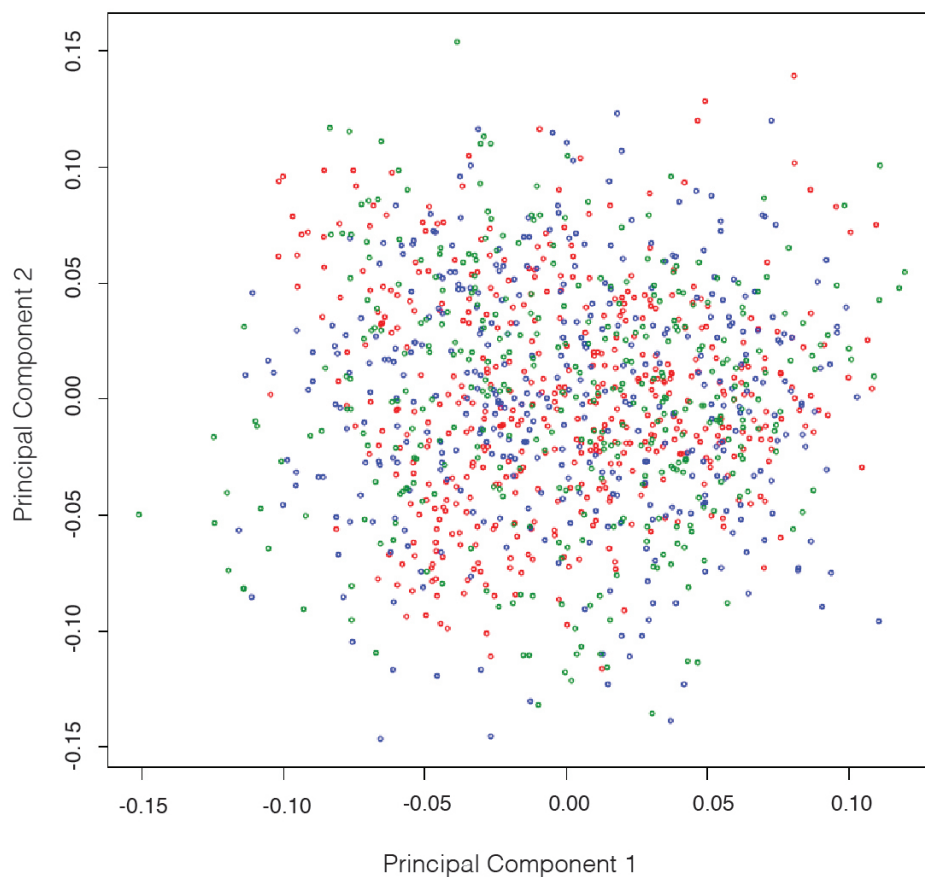
30. King, E. G. *et al.* Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome Res.* **22**, 1558–1566 (2012).
31. King, E. G., Macdonald, S. J. & Long, A. D. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics* **191**, 935–949 (2012).
32. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008).
33. Ackermann, M. & Beyer, A. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS Genet.* **8**, e1002463 (2012).
34. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
35. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* **9**, 855–867 (2008).
36. Cheverud, J. M. & Routman, E. J. Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455–1461 (1995).
37. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
38. Luner, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
40. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
41. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w¹¹¹⁸*; iso-2; iso-3. *Fly (Austin)* **6**, 80–82 (2012).
42. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
43. Kover, P. X. *et al.* A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000551 (2009).



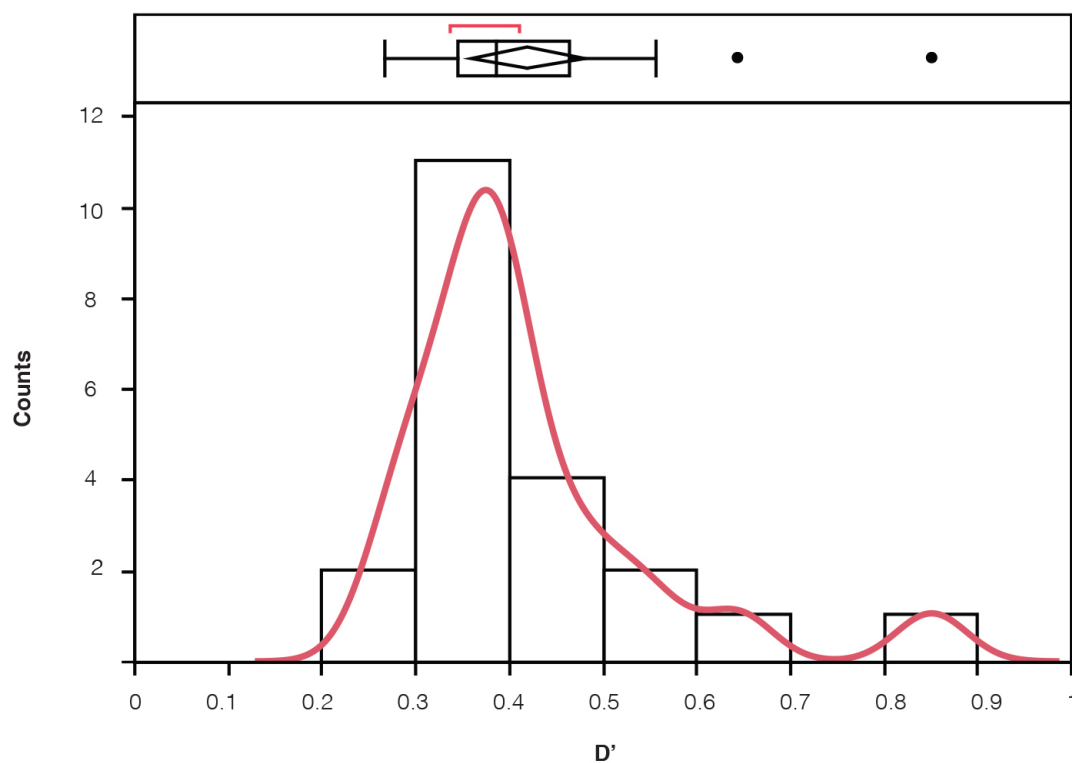
Extended Data Figure 1 | Description of the DSPR and validation scheme.

a. Geographic distribution of the DSPR founding strains (orange, panel A; red, panel B). **b.** Construction of the recombinant inbred lines. For each panel all founder strains were crossed in a round-robin design (line 1 ♀ × line 2 ♂, line 2 ♀ × line 3 ♂, ..., line 8 ♀ × line 1 ♂) to produce F₁s, and the F₁s were then allowed to mate free to produce an F₂ population. In each panel A and B, these F₂ populations were split into two independent population to create panels A1, A2 and B1, B2. Each was allowed to recombine freely for 50 generations, in very large population. After 50 generations, for each replicate

panel, about 400 isofemale lines were inbred for 25 generations to create the 4 panels of RILs used in this study. **c.** Crossing scheme used to validate epistatic effects. A pair of founder segregating incompatible alleles was selected and crossed to produce F₁s; we then intercrossed the F₁ progeny to produce a large F₂ population, segregating all possible allelic combinations between alleles at loci 1 and 2. We then counted the progeny each pair produced by intercrossing a large number of F₂s which were later genotyped at sites near to the predicted interacting loci.

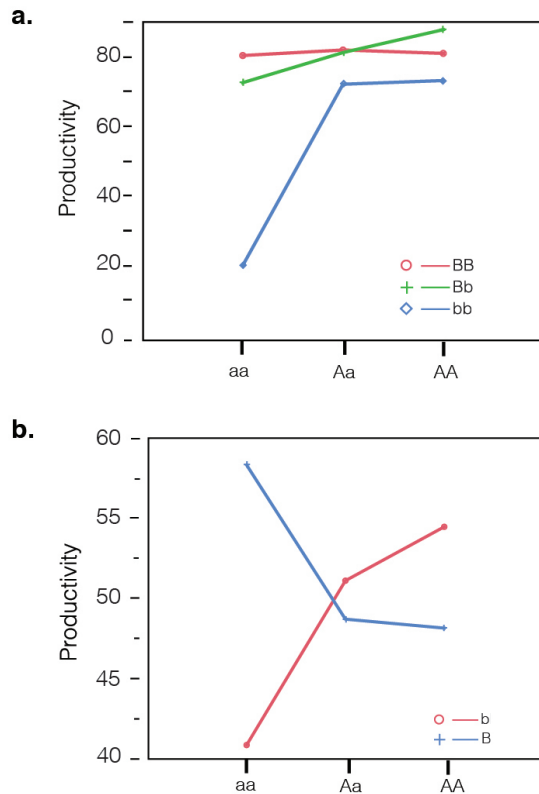


Extended Data Figure 2 | Principal component analysis of all three DSPR RIL panels. Green, panel A-2; blue, panel B-1; and red, panel B-2. No evidence of population structure is shown.



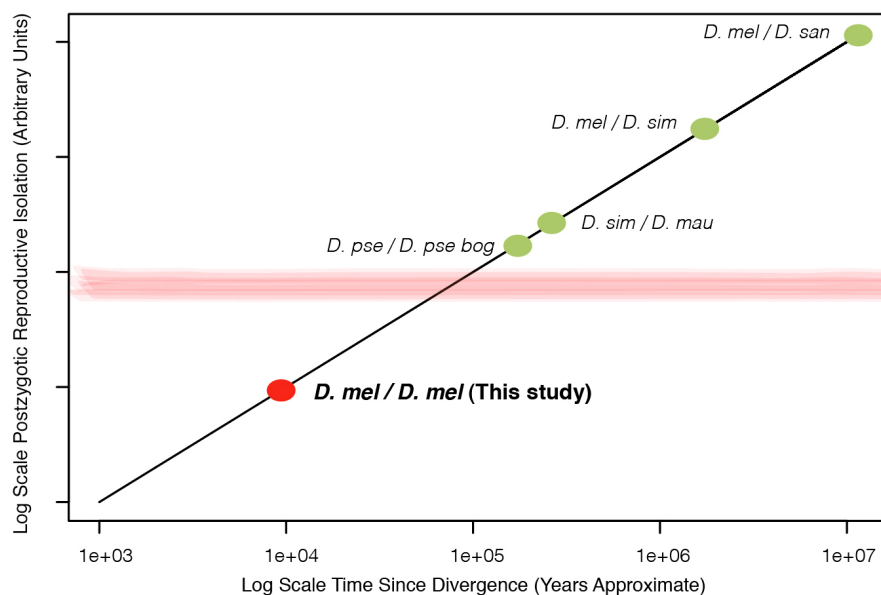
Extended Data Figure 3 | D' distribution for significant GRD. Data are plotted across DSPR panels. On the x axis, D' is a measure of the disequilibrium between interacting alleles. The red curve corresponds to a smooth curve fit using non-parametric density estimation. An outlier box-plot is presented

above the histogram (the lozenge represent the mean and 95% CI, the edge of the rectangle represent the 25% and 75% percentile, the vertical bar within the median, the dots are possible outlier and the red bracket represents the shortest length that contain 50% of the data).



Extended Data Figure 4 | Epistasis plot for each validated instance of GRD.

On the y axes are the productivity measurements that correspond to each genotypic class across both chromosomes. The x axes correspond to the genotypes on one of the chromosomes, the other genotype is represented by the colour indicated inside the plot (for example, genotype AA,bb in panel **a** is found in the lower left corner, where AA is read from the x axis and bb from the blue colour). **a**, GRD between chromosomes 2R and 3R (tagged by SNPs 2R:4806926, on the X axis and 3R:5870973, coloured lines) shows strong negative epistasis due to the low fitness of the $aa;bb$ genotype. The additive-by-additive genetic effect is equal to -13.75 (in the sense of refs 5 and 29). **b**, GRD between chromosomes 3L and X (tagged by SNPs 3L: 11510853, on the X axis and X: 16483812, coloured lines) also shows negative epistasis. Here the additive-by-additive genetic effect equals -5.94 .



Extended Data Figure 5 | The accumulation of post-zygotic reproductive isolation through time (note log scale on axes). Approximate divergence times of commonly studied *Drosophila* species are indicated by green circles, and the red circle indicates a reasonable expectation for divergence times of

stocks used to found the DSPR ($\sim 10,000$ years). The horizontal red area indicates a very approximate 'speciation threshold', and indicates that many species pairs that are commonly studied substantially exceed this threshold.

Extended Data Table 1 | List of all significant inter-chromosomal GRD identified in the DSPR

SNP based analysis																						
Panel	Chromosome 1	Position 1	Chromosome 2	Position 2	Number of RILs counted	1st Major Allele	1st Minor Allele	2dn Major Allele	2dn Minor Allele	1st Major allele freq	1st Minor allele freq	2dn Major allele freq	2dn Minor allele freq	Major-Major frequency	Major-Minor frequency	Minor-Major frequency	Minor-Minor frequency	d	d'	r	Chi-Square	p-value
B-2	2L	2767815	3R	4492436	391	C	A	A	T	0.91	0.09	0.92	0.08	0.86	0.05	0.06	0.03	0.02	0.27	0.25	23.70	5.85E-07
B-2	2L	8027605	X	13053319	418	C	T	C	T	0.94	0.06	0.88	0.12	0.85	0.10	0.03	0.03	0.02	0.38	0.25	26.89	1.12E-07
B-2	2L	10869984	3R	10633352	177	C	G	T	indel	0.94	0.06	0.95	0.05	0.92	0.03	0.03	0.02	0.02	0.41	0.39	26.77	1.18E-07
B-2	2L	21657908	3R	5870973	444	A	C	G	A	0.78	0.22	0.85	0.15	0.70	0.08	0.15	0.07	0.04	0.34	0.26	30.69	1.56E-08
B-2	2R	4806926	3R	5870973	443	G	A	G	A	0.50	0.50	0.85	0.15	0.49	0.01	0.36	0.14	0.06	0.85	0.36	58.05	1.30E-14
B-2	2R	8464341	X	5753834	457	A	G	T	A	0.89	0.11	0.77	0.23	0.72	0.17	0.05	0.06	0.03	0.39	0.25	29.29	3.22E-08
B-2	2R	20512785	X	19647595	436	T	G	A	C	0.64	0.36	0.59	0.41	0.33	0.32	0.27	0.09	-0.06	-0.40	-0.24	25.79	1.98E-07
B-2	3L	9627942	X	13622563	263	G	T	A	G	0.94	0.06	0.94	0.06	0.89	0.04	0.04	0.02	0.02	0.31	0.31	24.99	3.00E-07
B-2	3R	20437352	X	19127400	385	C	T	A	T	0.90	0.10	0.95	0.05	0.87	0.03	0.08	0.02	0.02	0.36	0.26	26.13	1.65E-07
B-1	2L	66907	3L	11787066	256	A	G	C	T	0.92	0.08	0.89	0.11	0.85	0.07	0.04	0.04	0.03	0.39	0.33	27.30	9.03E-08
B-1	2L	22311178	3R	5598460	375	A	G	A	T	0.75	0.25	0.90	0.10	0.71	0.04	0.19	0.06	0.04	0.48	0.28	28.94	3.85E-08
B-1	2L	2896002	X	11823233	274	A	G	T	G	0.95	0.05	0.93	0.07	0.90	0.05	0.03	0.02	0.02	0.36	0.31	26.88	1.12E-07
B-1	2L	10065007	3R	11588046	283	C	A	T	G	0.92	0.08	0.87	0.13	0.83	0.08	0.04	0.04	0.03	0.45	0.35	35.10	1.61E-09
B-1	2L	4140219	X	11514794	361	G	A	T	indel	0.94	0.06	0.95	0.05	0.91	0.03	0.04	0.02	0.02	0.33	0.30	33.13	4.43E-09
B-1	2R	3232234	3R	5598051	356	G	A	G	A	0.75	0.25	0.92	0.08	0.72	0.03	0.19	0.06	0.04	0.56	0.29	30.98	1.34E-08
B-1	2R	14543771	3R	22691609	355	A	G	T	C	0.85	0.15	0.93	0.07	0.81	0.04	0.12	0.04	0.03	0.41	0.27	26.33	1.49E-07
B-1	3R	13807981	X	8763898	151	T	C	C	T	0.91	0.09	0.88	0.12	0.84	0.07	0.04	0.05	0.04	0.51	0.45	30.06	2.17E-08
B-1	3R	18284739	X	14686047	378	G	T	G	A	0.94	0.06	0.92	0.08	0.88	0.06	0.04	0.02	0.02	0.30	0.25	23.46	6.62E-07
A-2	2L	19531958	X	12584624	326	G	T	C	G	0.94	0.06	0.94	0.06	0.91	0.04	0.03	0.02	0.02	0.35	0.34	37.94	3.74E-10
A-2	3L	11510853	X	16483812	354	A	T	G	A	0.92	0.08	0.91	0.09	0.86	0.06	0.05	0.03	0.02	0.28	0.27	26.16	1.62E-07
A-2	3R	23793328	X	14472525	64	A	T	C	T	0.84	0.16	0.84	0.16	0.80	0.05	0.05	0.11	0.08	0.64	0.64	26.58	1.31E-07
A-2	2L	16549805	3L	10566820	236	C	T	A	G	0.92	0.08	0.92	0.08	0.87	0.05	0.05	0.03	0.03	0.37	0.37	32.37	6.55E-09

HMM based analysis															
Panel	Chromosome 1	Position 1	Chromosome 2	Position 2	Number of RILs counted	1st Major Allele	1st Minor Allele	2dn Major Allele	2dn Minor Allele	Incompat. Panel-1 A	Incompat. Panel-1 B	Incompat. Panel-2 A	Incompat. Panel-2 B	Maximum Chi-square	Maximum P-value
B-2	2L	2767815	3R	4492436	391	C	A	A	T	2,3	6,7,8	1,5	2,3,4,6,7	22.285	1.22E-06
B-2	2L	8027605	X	13053319	418	C	T	C	T	2,7	1,3,6,8	5,6	1,2,3,7	28.552	4.71E-08
B-2	2L	10869984	3R	10633352	177	C	G	T	indel	1,3	6,7	2	3,4,5,6,7	14.007	9.69E-05
B-2	2L	21657908	3R	5870973	444	A	C	G	A	1,2,7	3,6,8	2	3,4,5,6,7	31.207	1.19E-08
B-2	2R	4806926	3R	5870973	443	G	A	G	A	3	1,2,6,7,8	2	3,4,5,6,7	62.734	1.20E-15
B-2	2R	8464341	X	5753834	457	A	G	T	A	1,3,8	2,5,6	1,4,5	2,3,7	37.070	5.84E-10
B-2	2R	20512785	X	19647595	436	T	G	A	C	4,8	2,3,5,6,7	3,5	1,2,4,6	28.937	3.86E-08
B-2	3L	9627942	X	13622563	263	G	T	A	G	2,3,5	1,6,7,8	5,7	1,3,6	17.625	1.41E-05
B-2	3R	20437352	X	19127400	385	C	T	A	T	3,5	1,4,6	8	2,4,5,6,7	10.122	7.95E-04
B-1	2L	66907	3L	11787066	256	A	G	C	T	7	1,2,3,6,8	5	2,3,4,6,7	16.936	2.04E-05
B-1	2L	22131178	3R	5598460	375	A	G	A	T	1,2,3,6	4,7,8	1,2,3,6	4,7,8	28.154	5.79E-08
B-1	2L	2896002	X	11823233	274	A	G	T	G	6,7,8	2,3,4,5	2,8	1,3,5,6,7	27.374	8.67E-08
B-1	2L	10065007	3R	11588046	283	C	A	T	G	1,7	2,3,4,6	2,8	1,2,3,6,7	29.341	3.13E-08
B-1	2L	4140219	X	11514794	361	G	A	T	indel	3	2,4,6,7,8	3,5,6	1,2,7,8	44.297	1.44E-11
B-1	2R	3232234	3R	5598051	356	G	A	G	A	4,7,8	1,2,3,6	4,7,8	1,2,3,6	30.049	2.17E-08
B-1	2R	14543771	3R	22691609	355	A	G	T	C	2,6	1,3,5,8	4	2,3,5,6,7	26.540	1.34E-07
B-1	3R	13807981	X	8763898	151	T	C	C	T	3,4	1,2,6,7,8	5,7,8	1,3,6	18.243	1.02E-05
B-1	3R	18284739	X	14686047	378	G	T	G	A	2,7	1,4,6,8	5	1,3,6,8	29.620	2.71E-08
A-2	2L	19531958	X	12584624	326	G	T	C	G	1	2,4,5,7	5	2,3,4,6,7	42.031	4.59E-11
A-2	3L	11510853	X	16483812	354	A	T	G	A	4	2,3,5,6	5	1,3,4,6,7	34.180	2.58E-09
A-2	3R	23793328	X	14472525	64	A	T	C	T	5	3,4,6,7	1,5	2,3,4,6,7	12.913	1.74E-04
A-2	2L	16549805	3L	10566820	236	C	T	A	G	1,8	2,4,5,7	4	3,5	20.660	2.86E-06

Extended Data Table 2 | List of significant inter-chromosomal GRD in the *Arabidopsis* MAGIC panel and maize NAM panel

MAGIC panel

Chromosome 1	Position 1	Chromosome 2	Position 2	Number of RIL counted	Chi Square	P-value
4	15897262	5	25151311	522	53.75	2.27E-13
1	4947328	4	11984761	513	47.07	6.85E-12
1	5447591	4	11984761	513	45.49	1.54E-11
2	5666979	3	15317766	442	41.02	1.51E-10
3	8190257	5	23705451	524	32.09	1.47E-08
1	30044327	3	21116640	517	24.91	6.00E-07
1	2211127	4	3781192	522	19.58	9.64E-06

NAM panel

PANEL	Marker 1 ID	Chromosome 1	Position 1	Marker 2 ID	Chromosome 2	Position 2	Number of RIL counted	Chi Square	P-value
4	PZB01647.1	1	137.6	PZA01951.1	8	32.3	175	27.50	1.57E-07
25	PZA03032.19	3	82.1	PHM7898.10	7	111.8	164	25.00	5.73E-07
12	PZA00494.2	3	97.8	PZA01964.29	8	106.9	136	24.88	6.09E-07
7	PZA01753.1	2	40.6	PZB00752.1	7	71.2	175	24.70	6.72E-07
4	PZA00261.6	5	66.8	PZA00090.1	8	70.6	160	23.58	1.20E-06

A melanocyte lineage program confers resistance to MAP kinase pathway inhibition

Cory M. Johannessen^{1,2,3}, Laura A. Johnson^{1,2†}, Federica Piccioni¹, Aisha Townes¹, Dennie T. Frederick⁴, Melanie K. Donahue¹, Rajiv Narayan¹, Keith T. Flaherty⁴, Jennifer A. Wargo⁴, David E. Root¹ & Levi A. Garraway^{1,2,3}

Malignant melanomas harbouring point mutations (Val600Glu) in the serine/threonine-protein kinase BRAF (BRAF(V600E)) depend on RAF–MEK–ERK signalling for tumour cell growth¹. RAF and MEK inhibitors show remarkable clinical efficacy in BRAF(V600E) melanoma^{2,3}; however, resistance to these agents remains a formidable challenge^{2,4}. Global characterization of resistance mechanisms may inform the development of more effective therapeutic combinations. Here we carried out systematic gain-of-function resistance studies by expressing more than 15,500 genes individually in a BRAF(V600E) melanoma cell line treated with RAF, MEK, ERK or combined RAF–MEK inhibitors. These studies revealed a cyclic-AMP-dependent melanocytic signalling network not previously associated with drug resistance, including G-protein-coupled receptors, adenylyl cyclase, protein kinase A and cAMP response element binding protein (CREB). Preliminary analysis of biopsies from BRAF(V600E) melanoma patients revealed that phosphorylated (active) CREB was suppressed by RAF–MEK inhibition but restored in relapsing tumours. Expression of transcription factors activated downstream of MAP kinase and cAMP pathways also conferred resistance, including *c-FOS*, *NR4A1*, *NR4A2* and *MITF*. Combined treatment with MAPK-pathway and histone-deacetylase inhibitors suppressed MITF expression and cAMP-mediated resistance. Collectively, these data suggest that oncogenic dysregulation of a melanocyte lineage dependency can cause resistance to RAF–MEK–ERK inhibition, which may be overcome by combining signalling- and chromatin-directed therapeutics.

To identify genes whose upregulation confers resistance to MAPK pathway inhibition, we expressed 15,906 human open reading frames (ORFs)⁵ (Extended Data Fig. 1) in a BRAF(V600E)-mutant, MAPK-pathway-dependent melanoma cell line (A375)^{6,7} and determined their effects on sensitivity to small-molecule inhibitors targeting RAF, MEK, ERK⁸ or a combination of RAF and MEK (Fig. 1a). In this experiment, 14,457 genes (90.9%, Fig. 1a) passed quality-control filters and were evaluated for their effects on drug sensitivity (Extended Data Fig. 2a, b, c). We identified 169 genes (1.16%) whose overexpression conferred resistance to at least one MAPK-pathway inhibitor (Extended Data Fig. 2d).

These screens identified diverse resistance effectors (Fig. 1b), including genes that activate ERK signalling (*KRAS*(Gly12Val), *MEK1*(Ser218/222Asp), *RAF1*, *MOS*, *FGR*, *AXL*, *FGFR2*, *SRC* and *MAP3K8* (also known by its protein abbreviation COT))^{6,9–13} and RAS–guanine exchange factors (*RASGRP2*, *RASGRP3* and *RASGRP4*) (Extended Data Fig. 2d). Previously unrecognized resistance mechanisms were also identified, including modifiers of ‘stem-ness’ (*OCT4* (also known as *POU5F1*), *NANOG*), ubiquitin pathway components (*KLHL*-family members, *TRIM*-family members) and non-Ras guanine exchange factors (*VAV1*, other *DBS* and *PLEKHG* family members). Furthermore, several ERK-regulated transcription factors emerged, including *FOS*, *JUNB*, *ETS2* and *ETV1* (Extended Data Fig. 2d).

To verify resistance effects, we re-expressed each candidate gene in A375 cells and calculated the area under the curve (AUC, Extended Data Fig. 3b) for MAPK-inhibitor growth inhibition (GI₅₀; concentration that inhibits growth by 50%) assays (Extended Data Fig. 3a). The fraction of candidate genes that were validated ($P < 0.05$) by these experiments ranged from 64.2% (RAF inhibitors) to 84.5% (RAF–MEK inhibitors) (Fig. 2a). Of the 75 RAF-inhibitor resistance genes, 71 (94.6%) also imparted resistance to MEK inhibitors and RAF–MEK inhibitors and only 18 (25.4%) of the 71 RAF-, MEK- and RAF–MEK-inhibitor resistance genes retained sensitivity to ERK inhibitors (Extended Data Fig. 3d, e). Thus, the majority of the genes that confer resistance to single agent RAF inhibitors are resistant to both RAF–MEK inhibitors (94.6%) and ERK inhibitors (70.6%) (Extended Data Fig. 3e, f). Aside from a subset of MAPKs and tyrosine kinases, most genes produced only minimal phosphorylated-ERK rescue in the presence of MAPK inhibitors (Extended Data Fig. 3c), consistent with the high degree of ERK-inhibitor resistance observed in our validation experiments (Fig. 2a). These data suggest that many resistance mechanisms may circumvent the entire RAF–MEK–ERK module.

We extended our validation studies across seven additional BRAF(V600E) lines (Extended Data Fig. 4a–d). Overall, 110 genes (66.7%) conferred resistance to the query inhibitors in at least two of seven additional BRAF(V600E) melanoma lines (Fig. 2b). Many genes again conferred resistance to all inhibitors or combinations of inhibitors examined (Fig. 2b). Next, we organized resistance genes into mechanistically related classes and identified those that exhibited the most extensive validation across our BRAF(V600E) cell lines (Fig. 2c). Based on these criteria, G-protein-coupled receptors (GPCRs) emerged as the top-ranked protein class (Extended Data Fig. 4e). Each validated GPCR conferred resistance to all MAPK inhibitors tested (Fig. 2b). Many GPCRs activate adenylyl cyclase, which converts ATP (ATP) to cAMP¹⁴, the primary target of which is protein kinase A (PKA). Consistent with these observations, the adenylyl cyclase gene *ADCY9* was also identified as a resistance effector (Extended Data Fig. 2d and Extended Data Fig. 4f, g) and the catalytic subunit of PKA α (*PRKACA*) had the highest composite rescue score within the serine/threonine kinase class (Fig. 2b, c). Both genes conferred resistance across all MAPK-pathway inhibitors examined (Fig. 2b and Extended Data Fig. 4f).

We therefore reasoned that a signalling network characterized by GPCR activation and induction of adenylyl cyclase–cAMP–PKA may induce resistance to MAPK inhibitors in melanoma. This predicted network resembles a growth-essential lineage pathway in primary melanocytes, which require GPCR-mediated cAMP signalling for growth *in vivo*¹⁵. To test this hypothesis, we determined whether cAMP-mediated signalling was sufficient to confer resistance to MAP kinase pathway inhibitors. Both cAMP and the adenylyl cyclase activator forskolin increased intracellular cAMP (Extended Data Fig. 5a) and conferred resistance to all MAPK-pathway inhibitors queried across a panel of

¹The Broad Institute of Harvard University and Massachusetts Institute of Technology, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ²Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. ³Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA. ⁴Department of Surgical Oncology, Medical Oncology and Dermatology, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA. [†]Present address: Graduate Program in Cell and Molecular Physiology, Sackler School of Graduate Biomedical Sciences, School of Medicine, Tufts University, 145 Harrison Avenue, Boston, Massachusetts 02111, USA.

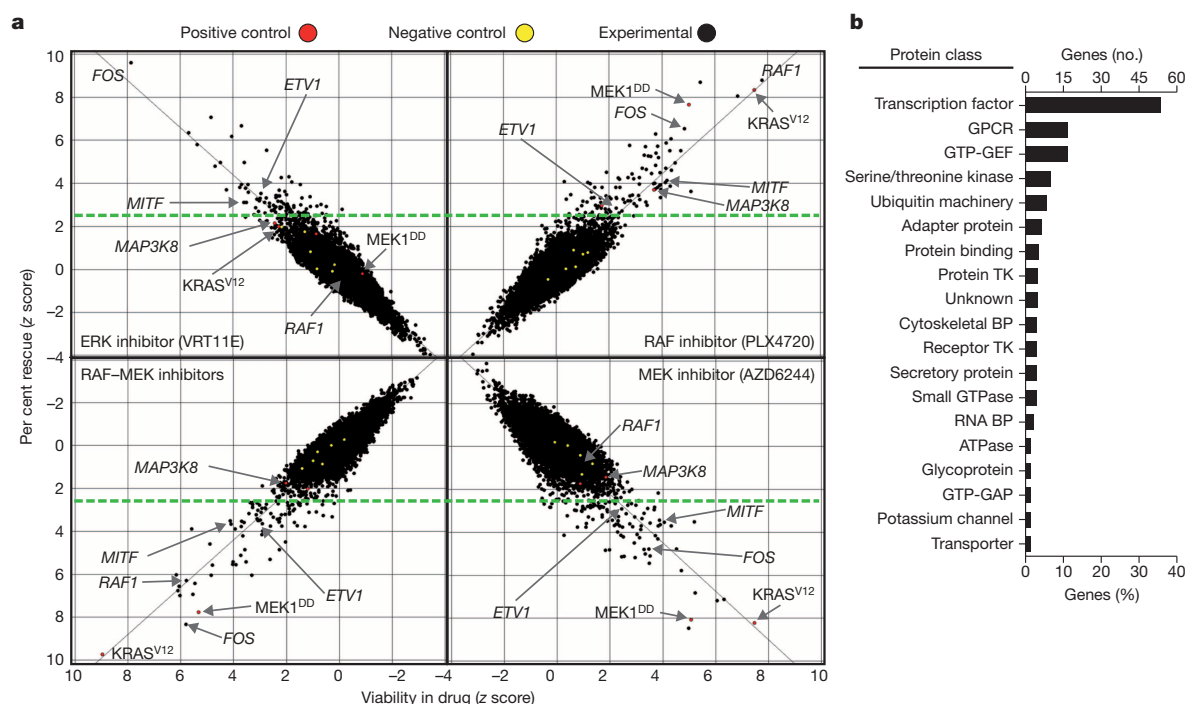


Figure 1 | Near-genome-scale functional rescue screens for resistance to RAF, MEK and ERK inhibitors. **a**, A375 cells transduced with the lentiviral expression library ($n = 2$ technical replicates) were treated with indicated inhibitors and assayed for viability in the presence of compound alone (x axis) and viability in compound relative to DMSO (y axis). Values are presented as a z score. Genes ($n = 169$) with z scores ≥ 2.5 (green dashed line) were nominated

as candidate resistance genes. **b**, Summary of protein classes of candidate genes identified in primary drug resistance screens. Only protein classes containing ≥ 2 genes are shown. GAP, GTP-activating protein; GEF, guanine nucleotide exchange factor; PTK, protein tyrosine kinase; RTK, receptor tyrosine kinase; TK, tyrosine kinase.

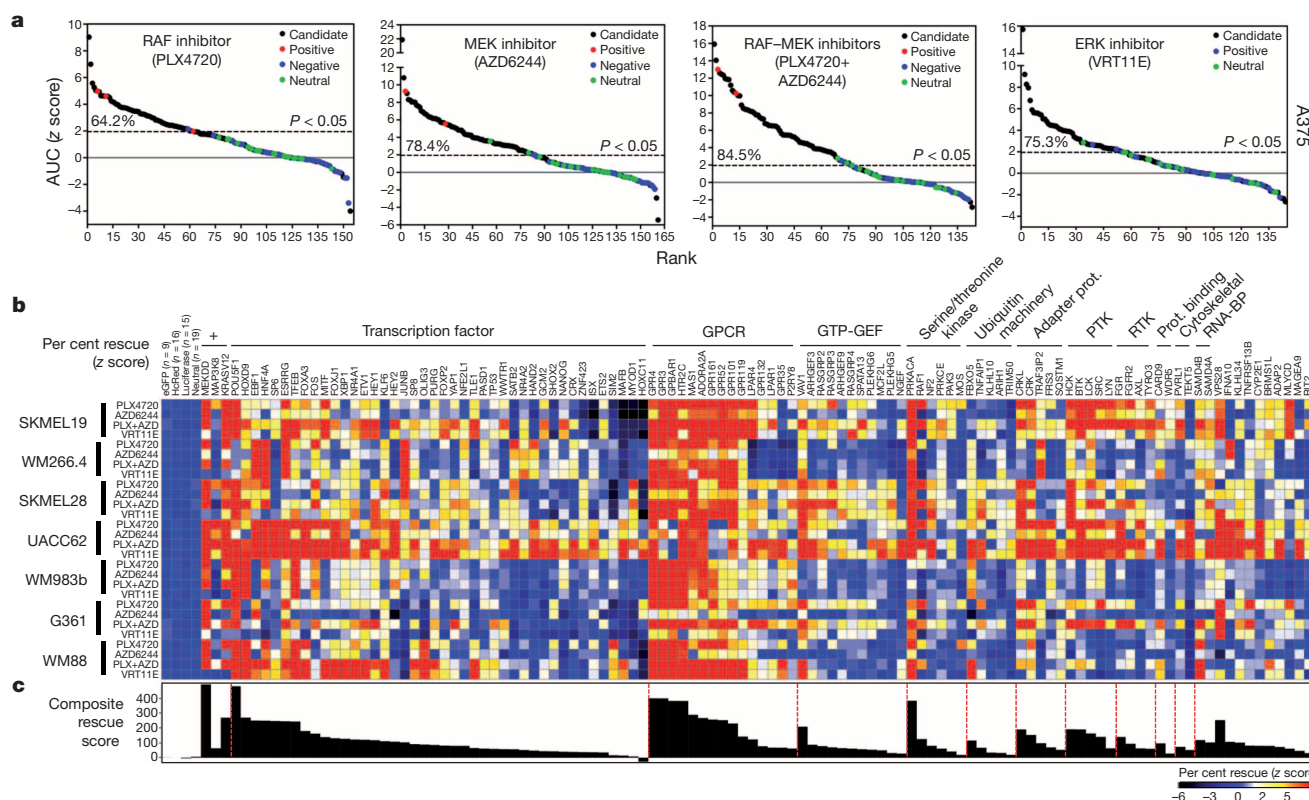


Figure 2 | Candidate resistance genes segregate into validating protein classes. **a**, Area under the curve (AUC) was calculated for MAPK-inhibitor drug sensitivity curves in A375 expressing candidate and control genes. Data are presented as a z score (y axis), relative to the AUC of all control genes across each MAPK inhibitor. **b**, Seven BRAF(V600E)-malignant melanoma cell lines expressing the indicated candidate or control genes were assayed ($n = 2$

technical replicates) for viability after treatment with indicated MAPK inhibitors. Cellular viability is presented as a z score relative to control genes. Genes with a z score ≥ 4 in ≥ 2 conditions (drug or cell line) are shown. **c**, The strength of the resistance phenotype for each candidate and control gene across all MAPK inhibitors and cell lines is quantified and presented as a composite rescue score.

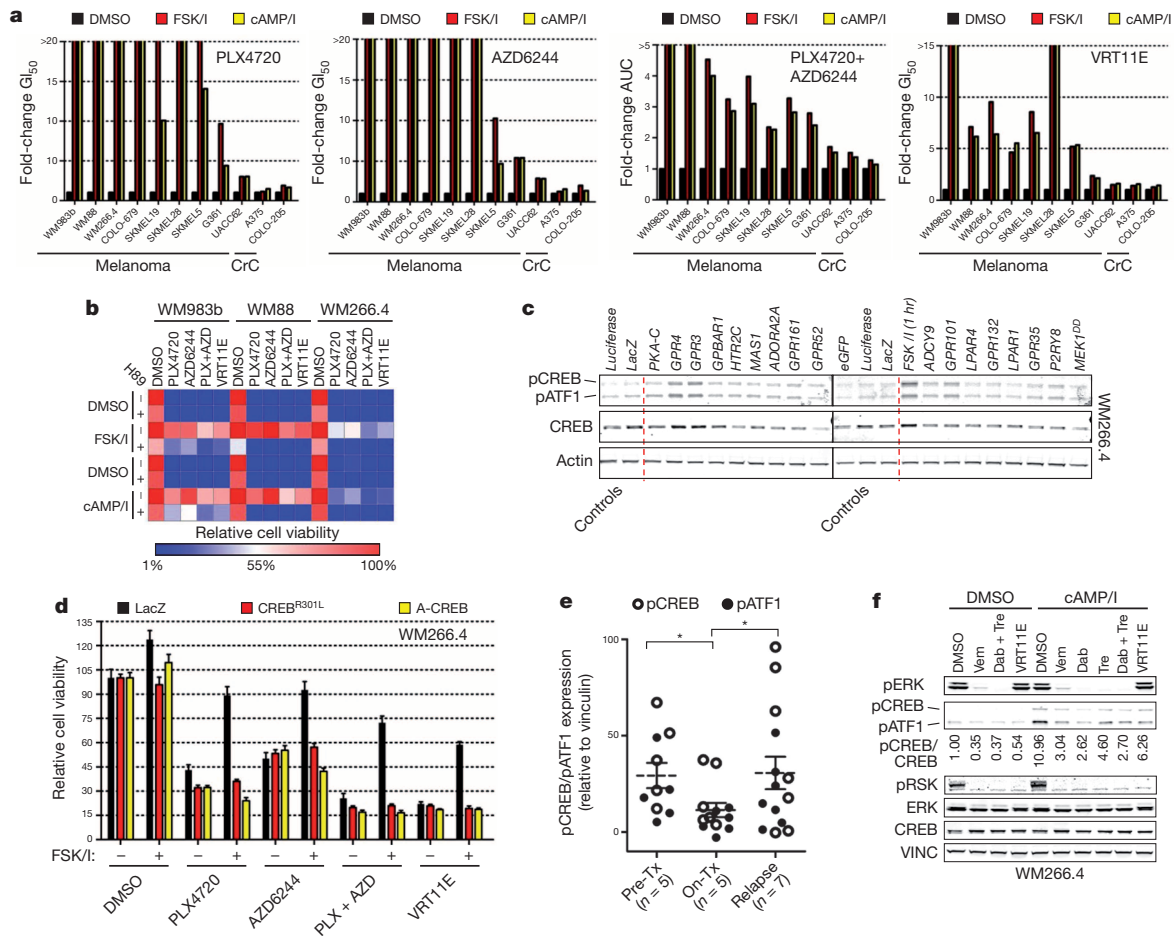


Figure 3 | A cyclic AMP signalling network mediates resistance to RAF, MEK and ERK inhibitors. **a**, Average fold change (relative to DMSO) in MAPK inhibitor G_{50} or AUC in a panel of BRAF(V600E)-mutant cell lines treated with vehicle (DMSO), forskolin and IBMX (FSK/I) or dibutyryl cAMP and IBMX (cAMP/I). $n = 8$ technical replicates, representative of 3 independent experiments. **b**, Heat map showing relative cell viability (per cent of DMSO) following treatment with FSK/I or cAMP/I in the presence of vehicle (DMSO), the PKA inhibitor H89 and a single dose of indicated MAPK inhibitors. **c**, Immunoblot analysis of phosphorylated CREB/ATF1 (Ser 133/Ser 63) in lysates from WM266.4 virally transduced with the indicated expression constructs. **d**, Viability of WM266.4 expressing either LacZ (control) or dominant-negative CREB alleles (CREB^{R301L} or A-CREB) following treatment with forskolin and IBMX (FSK/I) in the presence of

cell lines (Fig. 3a and Extended Data Fig. 5b) without affecting baseline cell growth (Extended Data Fig. 5c, d). Forskolin and cAMP resistance was PKA-dependent; it was blocked using the PKA inhibitor H89 (Fig. 3b and Extended Data Fig. 5e). The resistance phenotype was also relatively specific to MAPK-pathway inhibitors (Extended Data Fig. 5f). Thus, cAMP and PKA activation can confer resistance to MAPK-pathway inhibition in melanoma cells.

Two well-characterized transcription factor substrates of cAMP and PKA are CREB and ATF1, which regulate the expression of genes whose promoters harbour cyclic AMP response elements (CREs). To determine whether cAMP-mediated resistance may involve a CREB-dependent mechanism, we measured phosphorylation of these proteins following addition of either forskolin or exogenous cAMP. Both agents (Extended Data Fig. 6a, b), as well as most GPCR genes (Fig. 3c and Extended Data Fig. 6c, d), induced CREB and ATF1 phosphorylation, although only a subset of GPCRs increased steady-state intracellular cAMP (Extended Data Fig. 6e). Expression of dominant-negative CREB proteins (CREB^{R301L} (ref. 16) or A-CREB¹⁷; Extended Data Fig. 6f) suppressed forskolin-induced resistance to all MAPK-pathway inhibitors

indicated MAPK inhibitors. Viability is expressed as a percentage of DMSO. Error bars represent s.d. of mean, $n = 6$ technical replicates, representative of 2 independent experiments. **e**, Quantification of phosphorylated CREB (pCREB) and pATF1 expression following immunoblot analysis of lysates extracted from BRAF(V600E)-mutant human tumours. Tumours were biopsied pre-initiation of treatment ($n = 5$), following 10–14 days of MAPK-inhibitor treatment (on-treatment, $n = 6$) or following relapse ($n = 7$). MAPK-inhibitor therapy is noted. All available samples were tested and reported. Pre-treatment and on-treatment samples are paired. $*P < 0.05$, one-tailed t -test on treatment cohorts, which may not directly inform responses in individual patient samples. **f**, Immunoblot analysis of lysates from WM266.4 following treatment with FSK/I or cAMP/I in the presence of indicated MAPK inhibitors. Quantification of pCREB relative to CREB is shown.

tested (Fig. 3d). These results support the hypothesis that cAMP-mediated resistance may operate in large part through a CREB-dependent mechanism, though the roles of other downstream effectors cannot be excluded.

We next assessed the possible contribution of a cAMP–PKA–CREB mechanism in BRAF(V600E) melanoma patients by measuring CREB and ATF1 phosphorylation in tumour biopsies obtained before or during treatment and following relapse with vemurafenib alone or dabrafenib and trametinib in combination (Extended Data Fig. 7a). In contrast to cell lines *in vitro*, CREB and ATF1 phosphorylation was detectable in pre-treatment BRAF(V600E) melanoma biopsy specimens (Fig. 3e and Extended Data Fig. 7b). These results were consistent with the fact that cAMP pathway agonists are excluded from melanoma tissue culture media *in vitro*. Levels of phosphorylated CREB and ATF1 were suppressed in the cohort of patients treated with RAF or RAF–MEK inhibition (Fig. 3e and Extended Data Fig. 7b, c, d). In contrast, the levels of CREB and ATF1 phosphorylation observed in patient cohorts upon tumour relapse were statistically indistinguishable from those detected in the pre-treatment cohort (Fig. 3e). However,

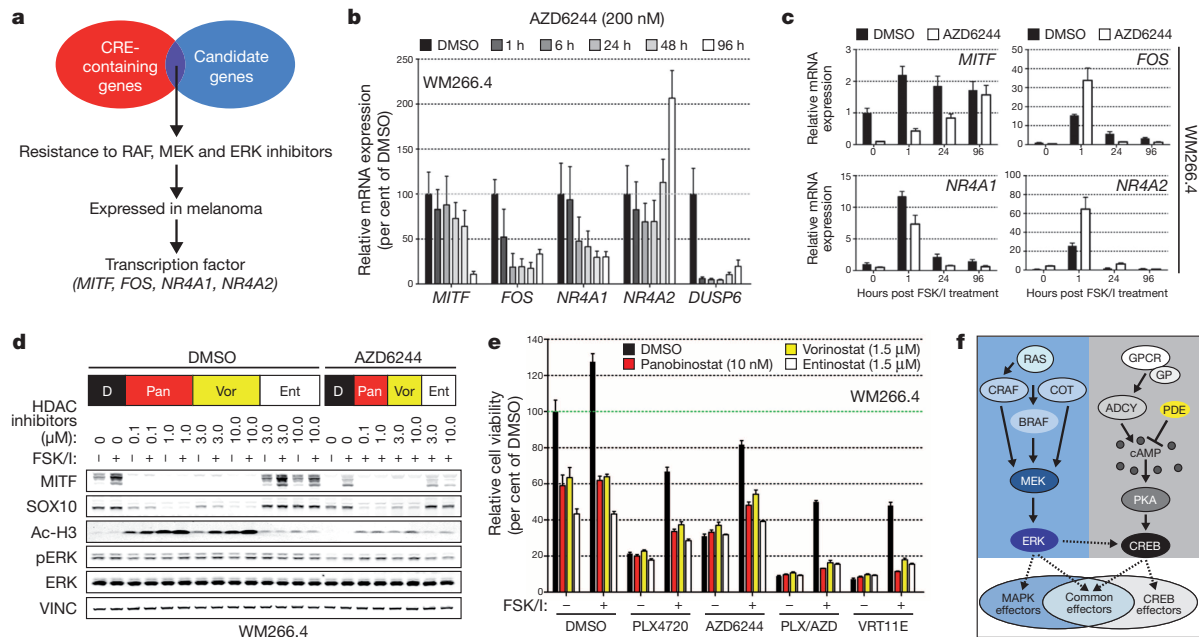


Figure 4 | Candidate resistance genes are transcriptional effectors of the MAPK and cAMP pathways. **a**, Schematic outlining the identification of candidate resistance genes endogenously regulated by cAMP. **b**, Quantification of TBP-normalized mRNA levels using real-time quantitative PCR (relative to DMSO treatment) following a time course of MEK inhibitor treatment. Error bars represent s.d. of mean, $n = 3$ technical replicates representative of 3 independent experiments. **c**, Quantification of TBP-normalized mRNA levels using real-time quantitative PCR (relative to DMSO-treatment) following treatment with forskolin and IBMX (FSK/I) for the indicated times in the presence of vehicle (DMSO) or MEK inhibitors. Error bars represent s.d. of

mean, $n = 3$ technical replicates representative of 2 independent experiments. **d**, Immunoblot analysis of lysates from WM266.4 treated with DMSO or MEK inhibitor followed by treatment with panobinostat, vorinostat or entinostat and subsequently stimulated with FSK/I. **e**, Cellular viability of WM266.4 treated with the indicated combinations of MAPK inhibitors, HDAC inhibitors and FSK/I. Cell viability is shown as a percentage of DMSO in un-stimulated or non-drug-treated cells. Error bars represent s.d. of mean, $n = 6$ technical replicates representative of 2 independent experiments. **f**, A mechanistic model of GPCR-mediated resistance. Solid arrows, direct regulation; dotted arrows, indirect regulation.

in the single case in which matched pre-, on-treatment and post-relapsed samples could be assessed, levels of phosphorylated CREB and ATF1 did not correlate with drug response (Extended Data Fig. 7b). These preliminary clinical results thus raise the possibility that a CREB-dependent mechanism might contribute to resistance to RAF-MEK inhibition in a subset of melanomas.

Based on these clinical findings, we sought to determine whether MAPK-pathway inhibitors might modulate levels of phosphorylated CREB and ATF1 *in vitro* when cAMP-dependent signalling is active. We treated BRAF(V600E) melanoma cells with cAMP and IBMX (a non-selective inhibitor) and measured phosphorylation of CREB and ATF1 following exposure to MAPK inhibitors. Each MAPK inhibitor partially blunted the increase in phosphorylated CREB and ATF1 produced by exogenous cAMP (Fig. 3f and Extended Data Fig. 7e), suggesting that cAMP-dependent activity of CREB and ATF1 may be reduced by pharmacologic MAPK inhibition.

In melanocytes, oncogenic BRAF or NRAS can substitute for cAMP signalling^{18–20}. We therefore reasoned that a cAMP-mediated lineage program might mediate resistance by inducing CREB-dependent transactivation of effectors normally under MAPK control (Fig. 4f). We identified CREs in the promoters of 19 resistance genes ($P = 5.0 \times 10^{-50}$; Fig. 4a and Extended Data Fig. 8a), of which three lineage-expressed (Extended Data Fig. 8c) transcription factors—MITF, FOS and NR4A2—showed high composite resistance scores ($z > 50$; Extended Data Fig. 8b). MITF, FOS, NR4A2 and NR4A1 (an NR4A2 homologue and validated resistance gene) showed reduced transcript levels following MEK inhibitor treatment (Fig. 4b). Activating MITF phosphorylation^{21,22} decreased within 1 h and total MITF protein was undetectable 48–96 h after MEK inhibition (Extended Data Fig. 9a, b). All four transcription factors exhibited 2- to 20-fold increases in messenger RNA expression within 1 h of forskolin treatment (Fig. 4c) and MITF showed sustained increases in protein

expression across multiple melanoma cell lines and MAPK pathway inhibitors (Extended Data Fig. 9c–f). Thus, CREB-responsive transcription factor resistance genes operate downstream of both MAPK- and cAMP-dependent signalling.

To further interrogate connections between cAMP signalling and resistance genes, we employed an expression profiling resource generated by the Library of Integrated Network-based Cellular Signatures (LINCS) program; an extensive catalogue of gene-expression profiles collected from human cells following chemical and genetic perturbation. We compared the signatures derived from all candidate resistance genes to a Library of Integrated Network-Based Cellular Signatures (LINCS) signature of adenylyl cyclase stimulation and found that the genes most similar to the signature of adenylyl cyclase activation were enriched for GPCR-pathway-associated candidate genes, including GPCRs, PKA and cAMP-MAPK-regulated transcription factors (Extended Data Fig. 9g). Thus, GPCR-pathway-related resistance genes and cAMP agonists function to elaborate a common transcriptional output.

Of the genes co-regulated by MAPK, and cAMP-CREB, MITF was intriguing because of its essential role in melanocyte development²³ and as a melanoma 'lineage survival' oncogene¹⁹. Expression of PKA α , ADCY9 or a subset of resistance-associated GPCRs enabled sustained MITF expression, even in the setting of MEK inhibitors (Extended Data Fig. 9h), thereby confirming that a GPCR-PKA-adenylyl cyclase cascade can regulate MITF expression in melanoma cells. Moreover, impairment of MITF protein levels by small hairpin RNA (shRNA) (Extended Data Fig. 10a, b) or co-treatment with a PKA inhibitor (H89, Extended Data Fig. 10c) blunted forskolin-mediated resistance to MAPK-pathway inhibitors (Fig. 3b and Extended Data Fig. 10a).

In a series of three patient-matched melanoma biopsies obtained over the course of RAF-MEK inhibition, we observed that MITF levels were reduced following initiation of MAPK-inhibitor therapy and partially restored in the context of relapse in one patient (Extended

Data Fig. 10d), consistent with the idea that aberrant expression of certain cAMP- and PKA-regulated transcription factors may correlate with resistance in some melanoma patients. Collectively, our findings indicate that resistance-associated transcriptional outputs may be governed by several transcription factors in melanoma cells.

Our results support a model in which aberrant signalling from melanocyte lineage pathways may converge on MITF or other transcription factors to drive resistance to MAPK pathway inhibitors. SOX10 and MITF expression can be impaired following treatment with histone deacetylase inhibitors (HDAC inhibitors), although these agents do not act exclusively through SOX10 and MITF²⁴. We reasoned that combined HDAC and MAPK inhibition might prevent cAMP- and MITF-driven resistance in melanoma cells. Indeed, multiple HDAC inhibitors (panobinostat (LBH589), vorinostat (SAHA) and entinostat (MS275)) reduced both SOX10 and MITF expression (Extended Data Fig. 10e), even in the presence of forskolin (Fig. 4d and Extended Data Fig. 10f). Each of these HDAC inhibitors reversed cAMP-mediated resistance to MAPK-pathway inhibition *in vitro* (Fig. 4e). Of note, forced expression of MITF did not abrogate HDAC-inhibitor sensitivity, indicating that the HDAC-inhibitor growth inhibitory effects do not act solely through this mechanism (Extended Data Fig. 10g). Nevertheless, these results raise the possibility that addition of HDAC inhibitors to combined RAF–MEK inhibition may offer a novel clinical strategy to achieve more durable control of some BRAF(V600E) melanomas.

The clinical benefit of RAF–MEK-inhibitor therapy in BRAF(V600E) melanoma remains temporary, and resistance mechanisms are incompletely understood. The GPCR–cAMP–adenyl cyclase–PKA–CREB module identified here is highly reminiscent of lineage survival signalling in melanocytes. Our results and those of other groups^{25,26} suggest that this lineage dependency may become reactivated as part of a clinical mechanism of resistance to RAF–MEK inhibition (Fig. 4f) and are bolstered by recent studies showing that MITF transcriptional targets are up regulated during the course of treatment with MAPK-pathway inhibitors²⁷. The application of genome-scale functional approaches to characterize anticancer drug resistance, together with directed experimental and clinical studies, may offer a general framework for discovery and clinical prioritization of novel therapeutic regimens.

METHODS SUMMARY

The arrayed ORF screens were performed as previously described⁶ using the Center for Cancer Systems Biology and Broad Institute Lentiviral Expression Library⁵. Effects of individual ORFs on drug sensitivity were determined by measuring differential viability (ratio of raw viability in MAPK-pathway inhibitor to viability in dimethylsulphoxide (DMSO)) and subsequently normalized across plates using a *z* score or standard score. Secondary screens to prioritize identified resistance candidates were performed in eight BRAF(V600E)-mutant melanoma cell lines in a manner similar to the primary screens. Prioritization of candidates was accomplished by generation of a composite rescue score for each gene, representing the extent and breadth of ORF-induced resistance phenotype across cell lines. Further validation and characterization of candidate resistance genes and pathways were accomplished using both biochemical and cell biological approaches. Detailed descriptions of all procedures are included in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 March; accepted 19 September 2013.

Published online 3 November 2013.

1. Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358–362 (2006).
2. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
3. Flaherty, K. T. *et al.* Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.* **367**, 1694–1703 (2012).

4. Flaherty, K. T. *et al.* Improved survival with MEK inhibition in BRAF-mutated melanoma. *N. Engl. J. Med.* **367**, 107–114 (2012).
5. Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nature Methods* **8**, 659–661 (2011).
6. Johannessen, C. M. *et al.* COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. *Nature* **468**, 968–972 (2010).
7. Wood, K. C. *et al.* MicroSCALE screening reveals genetic modifiers of therapeutic response in melanoma. *Sci. Signal.* **5**, rs4 (2012).
8. Aronov, A. M. *et al.* Structure-guided design of potent and selective pyrimidylpyrrole inhibitors of extracellular signal-regulated kinase (ERK) using conformational control. *J. Med. Chem.* **52**, 6362–6368 (2009).
9. Crews, C. M., Alessandrini, A. & Erikson, R. L. The primary structure of MEK, a protein kinase that phosphorylates the ERK gene product. *Science* **258**, 478–480 (1992).
10. Girotti, M. R. *et al.* Inhibiting EGF receptor or SRC family kinase signaling overcomes BRAF inhibitor resistance in melanoma. *Cancer Discov.* **3**, 158–167 (2012).
11. Kyriakis, J. M. *et al.* Raf-1 activates MAP kinase-kinase. *Nature* **358**, 417–421 (1992).
12. Patriotis, C., Makris, A., Chernoff, J. & Tsichlis, P. N. Tpl-2 acts in concert with Ras and Raf-1 to activate mitogen-activated protein kinase. *Proc. Natl Acad. Sci. USA* **91**, 9755–9759 (1994).
13. Pham, C. D., Arlinghaus, R. B., Zheng, C. F., Guan, K. L. & Singh, B. Characterization of MEK1 phosphorylation by the v-Mos protein. *Oncogene* **10**, 1683–1688 (1995).
14. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nature Rev. Mol. Cell Biol.* **3**, 639–650 (2002).
15. Hayward, N. K. Genetics of melanoma predisposition. *Oncogene* **22**, 3053–3062 (2003).
16. Walton, K. M., Rehfuess, R. P., Chrivia, J. C., Lochner, J. E. & Goodman, R. H. A dominant repressor of cyclic adenosine 3',5'-monophosphate (cAMP)-regulated enhancer-binding protein activity inhibits the cAMP-mediated induction of the somatostatin promoter *in vivo*. *Mol. Endocrinol.* **6**, 647–655 (1992).
17. Ahn, S. *et al.* A dominant-negative inhibitor of CREB reveals that it is a general mediator of stimulus-dependent transcription of c-fos. *Mol. Cell. Biol.* **18**, 967–977 (1998).
18. Dumaz, N. *et al.* In melanoma, RAS mutations are accompanied by switching signaling from BRAF to CRAF and disrupted cyclic AMP signaling. *Cancer Res.* **66**, 9483–9491 (2006).
19. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
20. Jané-Valbuena, J. *et al.* An oncogenic role for ETV1 in melanoma. *Cancer Res.* **70**, 2075–2084 (2010).
21. Hemesath, T. J., Price, E. R., Takemoto, C., Badalian, T. & Fisher, D. E. MAP kinase links the transcription factor Microphthalmia to c-Kit signalling in melanocytes. *Nature* **391**, 298–301 (1998).
22. Wu, M. *et al.* c-Kit triggers dual phosphorylations, which couple activation and degradation of the essential melanocyte factor Mi. *Genes Dev.* **14**, 301–312 (2000).
23. Hodgkinson, C. A. *et al.* Mutations at the mouse microphthalmia locus are associated with defects in a gene encoding a novel basic-helix-loop-helix-zipper protein. *Cell* **74**, 395–404 (1993).
24. Yokoyama, S. *et al.* Pharmacologic suppression of MITF expression via HDadenyl cyclase inhibitors in the melanocyte lineage. *Pigment Cell Melanoma Res.* **21**, 457–463 (2008).
25. Haq, R. *et al.* BCL2A1 is a lineage-specific antiapoptotic melanoma oncogene that confers resistance to BRAF inhibition. *Proc. Natl Acad. Sci. USA* **110**, 4321–4326 (2013).
26. Smith, M. P. *et al.* Effect of SMURF2 targeting on susceptibility to MEK inhibitors in melanoma. *J. Natl. Cancer Inst.* **105**, 33–46 (2013).
27. Frederick, D. T. *et al.* BRAF inhibition is associated with enhanced melanoma antigen expression and a more favorable tumor microenvironment in patients with metastatic melanoma. *Clinical Cancer Res.* **19**, 1225–1231 (2013).

Acknowledgements This work was supported by the National Institutes of Health (NIH) Director's New Innovator Award (DP2 OD002750, L.A.G.), Melanoma Research Alliance (L.A.G.), Starr Cancer Consortium (L.A.G.), Dr. Miriam and Sheldon G. Adelson Medical Research Foundation (L.A.G.), the NCI Skin Cancer SPORE (P50CA93683, L.A.G.) and the LINCS Program (U54 HG006093).

Author Contributions C.M.J. and L.A.G. designed the experiments. C.M.J. and L.A.J. performed primary and validation screens, with technical assistance from F.P. and supervision by D.E.R. All experimental follow-up studies were performed by C.M.J. A.T. performed quantitative PCR with reverse transcription (RT–PCR) experiments. M.K.D. generated gene signatures and R.N. analysed results. Clinical samples were collected or experiments performed by C.M.J., F.D.T., K.T.F., J.A.W. C.M.J. and L.A.G. wrote the manuscript. All authors discussed results and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.A.G. (Levi_Garraway@dfci.harvard.edu).

METHODS

Broad Institute and Center for Cancer Systems Biology Lentiviral Expression Library. The genesis, cloning, sequencing and production of the Broad Institute and Center for Cancer Systems Biology Lentiviral Expression Library have been described previously⁵. All ORFs described in this manuscript were expressed from pLX304 (<http://www.addgene.org/25890>), a lentiviral expression vector that encodes a C-terminal V5-epitope tag, a blasticidin resistance gene, and drives ORF expression from a cytomegalovirus (CMV) promoter. All clones described in this manuscript are publicly available through members of the ORFeome collaboration (<http://www.orfeomecollaboration.org/>).

Genome-scale ORF resistance screens. A375 cells were robotically seeded into 384-well white-walled, clear-bottom plates in RPMI-1640 (cellgro) supplemented with 10% FBS and 1% penicillin/streptomycin. The Broad Institute and Center for Cancer Systems Biology Lentiviral Expression Library⁵ was arrayed on 47 × 384-well plates, from which virus was robotically transferred to cell plates. Cell plates were randomly divided into six treatment arms in duplicate: DMSO, PLX4720, AZD6244, PLX4720 plus AZD6244, VRT11e or a parallel selection arm (blasticidin). Twenty-four hours after seeding, polybrene was added directly to cells ($7.5 \mu\text{g ml}^{-1}$ final concentration), followed immediately by robotic addition of the Broad Institute and Center for Cancer Systems Biology virus collection (3 μl per well) and centrifuged at 2250 r.p.m. (1,178g) for 30 min at 37 °C. Following a 24-h incubation at 37 °C (5% CO₂), media and virus was aspirated and replaced with complete growth media or media containing blasticidin ($10 \mu\text{g ml}^{-1}$) to select for ORF expressing cells and to determine infection efficiency. Forty-eight hours after media change, unselected (no blasticidin) cells were treated with DMSO (vehicle control) or MAPK pathway inhibitors to a final concentration of 2 μM (PLX4720, VRT11e) or 200 nM (AZD6244). Identical concentrations used for single-agent PLX4720 and AZD6244 treatment were used for combined PLX4720/AZD6244 treatment. Single-agent inhibitors were balanced with DMSO such that all wells contained 0.033% DMSO. Four days (96 h.) after drug addition, cell viability was assessed via robotic addition of CellTiterGlo (1:6 dilution) followed by 10 min orbital agitation at room temperature and subsequent quantification (EnVision Multilabel Reader, Perkin Elmer). Primary screens were performed in 16 individual batches in which two to three viral stock plates were screened per batch against all compounds.

Identification of resistance candidates from primary screening data. Following quantification of cell viability, duplicate luminescence values were averaged for each ORF within each treatment condition. Per cent rescue capability of each ORF was determined by dividing the average luminescence value in each drug by the average luminescence value in DMSO. Subsequent per cent rescue values were normalized within screening plates using the plate average and standard deviation to generate a *z* score or standard score of per cent rescue.

To calculate infection efficiency of each ORF, luminescence values in the presence of blasticidin were normalized to the average luminescence in DMSO and expressed as a percentage. ORF-mediated effects on cell viability in the absence of drug were assessed by taking the average luminescence value for each ORF in DMSO and normalizing each value to the plate average and standard deviation (*z* score).

To identify candidate resistance genes, we first filtered out all wells that had an infection efficiency of less than 65%. To eliminate genes with significant effects on cellular growth in the absence of drug treatment, we then filtered out genes that had a *z* score in DMSO of greater than 2.0 or less than -2.0. In addition, we eliminated from further analysis wells that showed a replicate variability (in DMSO) of greater than 29.15% (equivalent to >2 standard deviations from the average replicate variability). Following this initial filtering, 14,457 genes remained for subsequent analysis. Within each drug treatment condition, wells showing replicate variability of >2 standard deviations from the mean variability per drug were eliminated from further analysis. Finally, genes showing a *z* score of per cent rescue of >2.5 were nominated as resistance gene candidates.

Neutral control genes (19) were nominated from primary screening data by identifying genes across virus plates and screening batches with: high infection efficiency (>98.5%); minimal effects on baseline cell growth (*z* score of viability in DMSO between -0.5 to 0.5); and a rescue score (*z* score of per cent rescue) <0.25 (for example, no effect on drug sensitivity or resistance). DNA encoding candidates (169), negative controls (enhanced green fluorescent protein (eGFP), *n* = 9; HcRed, *n* = 15; luciferase, *n* = 16) positive controls (MEK1^{DD}, KRAS^{G12V}, MAP3K8/COT) and neutral controls (*n* = 19) were isolated from the Broad Institute and Center for Cancer Systems Biology expression collection and used to create a validation viral stock distinct from that used in the primary screens.

Drug sensitivity curves in A375 cells expressing candidate ORFs. A375 cells were seeded, infected and drug treated exactly as in primary screens using 4 μl of validation viral stock and concentrations of inhibitors ranging from 10 μM to 100 nM in half-log increments. For combinatorial PLX4720-AZD6244 treatment, a fixed dose of PLX4720 (2 μM) was combined with AZD6244 in doses ranging

from 10 μM to 100 nM in half-log increments. Viability was assessed as in the primary screen. Resulting luminescence for each ORF was normalized to luminescence in DMSO (per cent rescue) for each drug and drug concentration. Resulting sensitivity curves for each ORF were log transformed and the area under the curve (AUC) calculated using Prism GraphPad software. The resulting AUC for each candidate and control ORF-drug combination were normalized to that of the negative and neutral controls using a *z* score (described above). ORFs yielding a *z* score of >1.96 (*P* < 0.05) were considered to be validated candidates in this cell line.

Validation screens in additional BRAF(V600E) cell lines. Validation screening in additional BRAF(V600E) melanoma cell lines was performed exactly as in the primary screen, but cell lines were empirically optimized for seeding density and viral dilution. Owing to sensitivity of these cell lines to polybrene and virus exposure, all cell lines except for WM266.4 were treated with polybrene and virus, spun for 1 h at 2,250 r.p.m. (1,178g) followed immediately by complete virus and media removal and change to complete growth media. WM266.4 were treated with polybrene and virus, spun for 30 min at 2,250 RPM (1,178g) and incubated for 24 h before virus and media removal and change to complete growth media 24 h after infection. For experimental determination of infection efficiency, blasticidin ($5 \mu\text{g ml}^{-1}$) was added 24 h after media change. All drug treatments and viability measurements were performed as in primary screens.

The resulting luminescence values were normalized to DMSO (per cent of DMSO or 'per cent rescue'). The resulting per cent rescue was normalized to the mean and standard deviation of all negative and neutral controls to yield a *z* score of per cent rescue. Genes with a *z* score of per cent rescue of >4 in at least two instances were considered to have validated. 'composite rescue scores' were derived by summing the *z* score of per cent rescue of each gene across all drugs and cell lines. Average composite rescue scores for each protein class were generated by taking the average composite rescue score of all genes within a given protein class.

Phosphorylated ERK and V5 immunoassays. For analysis of ERK phosphorylation, A375 were seeded at 1,500 cells per well in black-walled, clear-bottomed, 384-well plates, virally transduced with all candidates and controls and treated with PLX4720, AZD6244 and combinatorial PLX4720-AZD6244 exactly as in the primary resistance screens. Eighteen hours after drug treatment, media was removed and cells were fixed with 4% formaldehyde and 0.1% Triton X-100 in PBS for 30 min at room temperature. Following removal of fixation solution, cells were washed once with PBS and blocked in blocking buffer (LiCOR) for 1 h at room temperature (21–25 °C) with shaking. After removal of blocking buffer, fixed cells were incubated with primary antibody against ERK phosphorylated at Thr 202/Tyr 204 (Sigma, #M8159, 1:2000) in LiCOR blocking buffer containing 0.1% Tween-20 and for 18 h at 4 °C with shaking. Antibody was removed and wells were washed thrice with 0.1% Tween-20 in water followed by incubation in secondary antibody (IRDye 800CW LiCOR, 1:1,200) and dual cellular stains, including Sapphire700 (LiCOR, 1:1000) and DRAQ5 (Cell Signaling Technology, 1:10,000), all diluted in LiCOR blocking buffer (no detergent) and incubated for 1 h at room temperature with shaking. Secondary antibody or cell stain was removed and washed thrice with 0.1% Tween-20 in water followed by a single wash in PBS. PBS was removed and plates were dried for 10 min at room temperature in the dark followed immediately by imaging on an Odyssey CLx Infrared Scanner. For phosphorylated ERK (pERK) and cellular stain, background was calculated based on signal observed in control wells containing only secondary antibody in blocking buffer and subtracted from each experimental well. Total pERK signal was normalized to total cellular stain for each ORF in each drug condition. The resulting values were subsequently normalized to DMSO (per cent of DMSO) for each ORF per drug condition.

V5 immunostaining for ectopic ORF expression was performed as described for the ERK phosphorylation assay above. In brief, cells were seeded at 3,000–4,000 cells per well and infected in parallel to validation screens. Seventy-two hours after infection, cells were fixed, blocked and stained as described for the pERK assay, instead using an antibody directed against the V5 epitope (Invitrogen, #R96025, 1:5,000, Invitrogen). Subsequent washes, secondary antibody incubations and total cellular staining protocol were identical to those described for the pERK assay, above. V5 and cellular stain (DRAQ5/Sapphire700) intensity were quantified as above, background signal subtracted (determined by signal intensity in uninfected wells with no V5 epitope and stained with secondary antibody, only) and V5 signal intensity normalized to cellular stain intensity.

Detection of GPCR-mediated cyclic AMP production. HEK293T cells were seeded at a density of 2.5×10^5 cells per well in 12-well plates. Twenty-four hours after seeding, cells were transfected with 250 ng of the indicated ORF (pLX304 expression vector) using 3 μl of Eugene6 (Promega) transfection reagent. Forty-seven hours after transfection, cells were treated either with DMSO (1:1,000) or IBMX (30 μM). In addition, forskolin (10 μM) and 100 M IBMX were added as

positive controls for indicated time. Cells were subsequently lysed in triton X-100 lysis buffer (Cell Signaling Technology) and resulting lysates split for cAMP enzyme-linked immunosorbent assay (ELISA) (Cell Signaling Technology, #4339) or parallel western blot analysis. cAMP ELISA was performed exactly as per the manufacturer's recommended protocol. Following quantification the inverse absorbance was calculated and normalized to that of negative control ORFs.

Identification of cyclic AMP response elements in candidate resistance genes.

Gene sets that share a common *CREB1*, *ATF1*, *ATF2* or *JUND* DNA response element within ± 2 kb of their transcriptional start site (as defined by TRANSFAC, version 7.4, <http://www.gene-regulation.com/>) were identified and downloaded from the MSigDB website (Extended Data Fig. 8a), available at <http://www.broadinstitute.org/gsea/msigdb>. CRE-containing genes present in individual gene sets were subsequently identified within the group of screened ORFs and within the group of candidate/neutral control ORFs. The ratio of CRE-containing genes to screened genes (expected) was compared to the ratio of CRE-containing genes to candidate/neutral control genes (actual) across gene sets. A *P* value for the observed enrichment of CRE-containing genes in the candidate genes over the expected representation within the screening set was calculated using Pearson's chi-squared test.

Cell lines and reagents. A375, SKMEL28, SKMEL19, UACC62, COLO-679 and WM983b cells were all grown in RPMI-1640 (Cellgro), 10% FBS and 1% penicillin and streptomycin. WM88, G361, SKMEL5, WM266.4, COLO-205 and 293T cells were all grown in DMEM (Cellgro), 10% FBS and 1% penicillin and streptomycin. All cell lines were acquired via the Cancer Cell Line Encyclopedia (<http://www.broadinstitute.org/ccle/home>), except for SKMEL19, which was a gift from N. Rosen. AZD6244 (PubChem ID: 10127622) was purchased from Selleck Chemicals, PLX4720 (PubChem ID: 24180719) was purchased from Symansis and VRT11e was synthesized by contract based on its published structure⁸. Forskolin, IBMX (3-Isobutyl-1-methylxanthine), cyclic AMP (cAMP, N6, 2'-*O*-dibutyryl adenosine 3':5'-cyclic monophosphate) and α -MSH (α -melanocyte stimulating hormone) were purchased from Sigma. Panobinostat (LBH-589) was purchased from BioVision, Vorinostat (SAHA) and Entinostat (MS-275) from were purchased from Cayman Chemical. All small molecules were dissolved in DMSO.

Pharmacologic growth inhibition assays. Melanoma cell lines were seeded into 384-well, white-walled, clear-bottom plates at the following densities; A375, 500 cells per well; SKMEL19, 1,500 cells per well; SKMEL28, 1,000 cells per well; UACC62, 1,000 cells per well; WM266.4, 1,800 cells per well; G361, 1,200 cells per well; COLO-679, 2,000 cells per well; SKMEL5, 2,000 cells per well; WM983b, 1,500 cells per well; WM88, 1,800 cells per well; COLO-205, 1,500 cells per well. Twenty-four hours after seeding, serial dilutions of the relevant compound were prepared in DMSO to 1000 \times stocks. Drug stocks were then diluted 1:100 into appropriate growth media and added to cells at a dilution of 1:10 (1 \times final), yielding drug concentrations ranging from 100 μ M to 1×10^{-5} μ M, with the final volume of DMSO not exceeding 1%. When indicated, forskolin (10 μ M), IBMX (100 μ M), dibutyryl cAMP (100 μ M) were added concurrent with MAPK-pathway inhibitors. Cells were incubated for 96 h following addition of drug. Cell viability was measured using CellTiterGlo viability assay (Promega). Viability was calculated as a percentage of control (DMSO treated cells). A minimum of six replicates were performed for each cell line and drug combination. Data from growth-inhibition assays were modelled using a nonlinear regression curve fit with a sigmoid dose-response. These curves were displayed and GI_{50} generated using GraphPad Prism 5 for Windows (GraphPad). Sigmoid-response curves that crossed the 50% inhibition point at or above 1.0 μ M or 10.0 μ M have GI_{50} values annotated as >1.0 μ M or >10.0 μ M, respectively. For single-dose studies, WM266.4 were seeded at 5,000 cells per well in 96-well, white-walled, clear-bottom plates and the identical protocol (above) was followed, using a single dose of indicated drug.

ORF and short hairpin RNA expression methods for experimental studies.

Indicated ORFs were expressed from pLX-304 (Blast, V5) lentiviral expression plasmids, whereas shRNAs were expressed from pLKO.1. shRNAs and controls are available through the RNAi Consortium Portal (<http://www.broadinstitute.org/rnai/public/>) and are identifiable by their sequence and clone ID: shLuc (CTT CGAAATGTCGTTCCGTT, TRCN0000072243), shMITF₄₉₂ (TTAGCCTA GAATCAAGTTATA, TRCN0000329869) and shMITF₅₇₃ (CGGGAAACTT GATTGATCTTT, TRCN0000019123). For lentiviral production, 293T cells (1.0×10^6 cells per 6-cm dish) were transfected with 1 μ g of pLX-Blast-V5-ORF or pLKO.1-shRNA, 900 ng Δ 8.9 (gag, pol) and 100 ng VSV-G using 6 μ l Fugene6 transfection reagent (Promega). Viral supernatant was collected 72 h post transfection. WM266.4 were infected at a 1:10–1:20 dilution (ORFs) or 1:100 dilution (shRNA) of virus in 6-well plates (2.0×10^5 cells per well, for immunoblot assays) or 96-well plates (3.0×10^3 , for cell growth assays) in the presence of 5.5 μ g ml⁻¹ polybrene and centrifuged at 2,250 r.p.m. for 60 min at 37 °C followed immediately by removal of media and replacement with complete growth media. Seventy-two hours after infection, drug treatments/pharmacological perturbations were initiated (see below).

Generation of CREB1 and A-CREB reagents. Wild-type CREB1 (Isoform B, NM_134442.3) was obtained through the Broad Institute RNAi Consortium, a member of the ORFeome Collaboration (<http://www.orfeomecollaboration.org/>). Arginine 301 of CREB was mutated to Leucine yielding CREB(R301L) (equivalent to CREB(R287L) in isoform A) using the QuikChange Lightning Mutagenesis Kit (Agilent), performed in pDonor223 (Invitrogen). CREB(R301L) was transferred into pLX304 using LR Clonase (Invitrogen) per manufacturer's recommendation. The A-CREB complementary DNA¹⁷ was synthesized (Genewiz) with flanking Gateway recombination sequences, recombined first into pDonor223 and subsequently into pLX304 as described for CREB1 mutant cDNAs.

Quantitative RT/PCR. mRNA was extracted from WM266.4 using the RNeasy kit (Qiagen) and homogenized using the Qiashredder kit (Qiagen). Total mRNA was used for subsequent reverse transcription using the SuperScript III First-Strand Synthesis SuperMix (Invitrogen). Five microlitres of reverse-transcribed cDNA was used for quantitative PCR using SYBR Green PCR Master Mix and gene-specific primers, in quadruplicate, using an ABI PRISM 7900 Real Time PCR System. Primers used for detection were as follows; NR4A2 forward: 5'- GTT CAG GCG CAG TAT GGG TC -3'; NR4A2 reverse: 5'- AGA GTG GTA ACT GTA GCT CTG AG -3'; NR4A1 forward: 5'- ATG CCC TGT ATC CAA GCC C -3'; NR4A1 reverse: 5'- GTG TAG CCG TCC ATG AAG GT -3'; DUSP6 forward: 5'- CTG CCG GGC GTT CTA CCT -3'; DUSP6 reverse: 5'- CCA GCC AAG CAA TGT ACC AAG -3'; MITF forward: 5'- TGC CCA GGC ATG AACACA C-3'; MITF reverse: 5'- TGG GAA AAA TACACG CTG TGA G -3'; FOS forward: 5'- CAC TCC AAG CGG AGA CAG AC -3'; FOS reverse: 5'- AGG TCA TCA GGG ATC TTG CAG -3'; TBP forward: 5'- CCC GAA ACG CCG AAT ATA ATCC C-3'; TBP reverse: 5'- GAC TGT TCT TCA CTC TTG GCT C -3'. Relative expression was determined using the comparative CT method (Applied Biosystems) followed by normalization to the DMSO/*T*₀ time point.

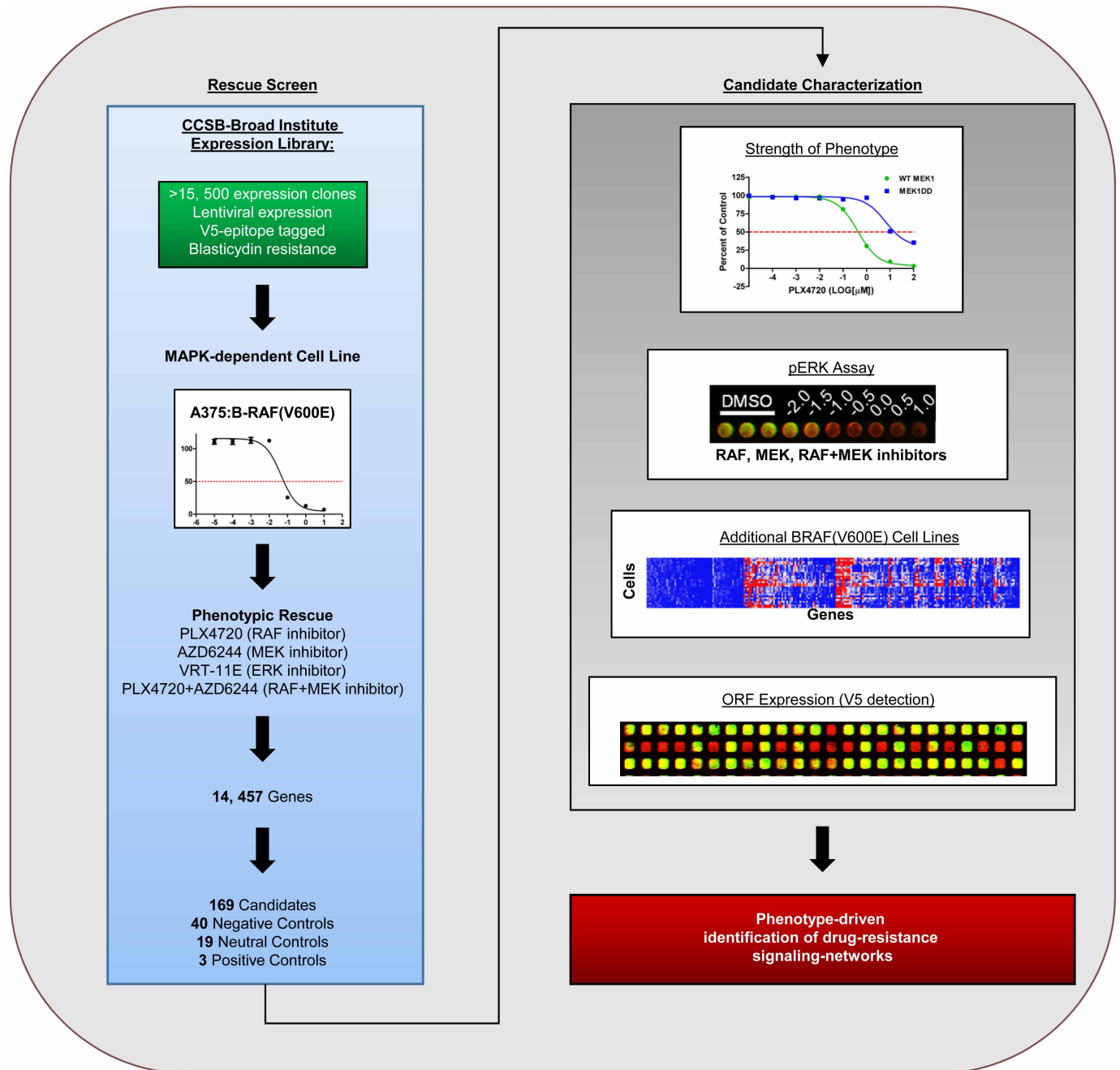
Immunoblot analyses and antibodies. Adherent cells were washed once with ice-cold PBS and lysed passively with 1% NP-40 buffer (150 mM NaCl, 50 mM Tris, pH 7.5, 2 mM EDTA, pH 8, 25 mM NaF and 1% NP-40) containing 2 \times protease inhibitors (Roche) and 1 \times Phosphatase Inhibitor Cocktails I and II (CalBioChem). Lysates were quantified (Bradford assay), normalized, reduced, denatured (95 °C) and resolved by SDS gel electrophoresis on 4–20% Tris or Glycine gels (Invitrogen). Resolved protein was transferred to nitrocellulose or PVDF membranes, blocked in LiCOR blocking buffer and probed with primary antibodies recognizing MITF (NeoMarkers, Clone C5, #MS-771-P, 1:400), Cyclin D1 (NeoMarkers, Clone Ab-3, #RB-010-P, 1:400), pERK1 and pERK2 (Thr 202/Tyr 204, Sigma, #M8159, 1:5,000), SLVR (Sigma, SAB4100050, 1:500), vinculin (Sigma, V9131, 1:20,000), total MEK1 (BD Transduction, #610122, 1:1000), acetylated histone H3 (Millipore, #06-599, 1:2000) and V5 epitope (Invitrogen, #R96025, 1:5,000). The following antibodies were purchased from Cell Signaling Technology and used at 1:1000 dilution: pMEK1 and pMEK2 (Ser 217/Ser 221, #9154), FOS (#2250), pCREB and pATF1 (Ser 133, Ser 63, respectively, #9196), CREB (#4820) and β -Actin (#3700, 1:20,000). The following antibodies were purchased from Santa Cruz Biotechnology: BCL2 (Clone C-2, sc-7382, 1:250), TRP1 (Clone G-17, sc-10443, 1:1000), Melan-A (Clone A103, sc-20032, 1:1000), ERK2 (Clone C-14, sc-154, 1:5,000), NR4A1/Nur77 (Clone M-210, sc-5569, 1:250), NR4A2/Nurr1 (Clone N-20, sc-991, 1:500), SOX10 (Clone N-20, sc-17342, 1:400). After incubation with the appropriate secondary antibody (anti-rabbit, anti-mouse or anti-goat immunoglobulin G (IgG), IRDye-linked; 1:15,000 dilution; IRDye 800CW, 1:20,000 IRDye 680LT, LiCOR), proteins were imaged and quantified using an Odyssey CLx scanner (LiCOR). Lysates from tumour and matched normal skin were generated by homogenization of tissue in 1% Triton X-100, 50 mM HEPES, pH 7.4, 150 mM NaCl, 1.5 mM MgCl₂, 1 mM EGTA, 100 mM NaF, 10 mM Na pyrophosphate, 1 mM Na₃VO₄, 10% glycerol, containing freshly added protease and phosphatase inhibitors (Roche Applied Science Cat. # 05056489001 and 04906837001, respectively). Subsequent normalization and immunoblot analyses were performed as above.

LINCS analysis. To explore transcriptional connections between cAMP signalling and GPCR-pathway-associated drug-resistance candidates, we expressed all of our candidate and control genes in A375 cells (as described above) and generated gene-expression profiles using a high-throughput Luminex bead-based platform. We queried the LINCS database (<http://www.lincscloud.org>) using a gene-expression signature of adenylyl cyclase stimulation generated by treating A375 cells with colforsin, an adenylyl cyclase agonist. We computed the similarity of the colforsin signature to 8729 treatment signatures in the A375 cell line (including the resistance candidate genes) that were available in the database, using a two-tailed weighted enrichment metric (connectivity score). We obtained a ranked list of the treatments based on the strength of the connectivity scores, and examined the ranks of the resistance candidate genes as well as the ranks of neutral control genes.

Expression profiling of melanoma cancer cell lines. We carried out oligonucleotide microarray analysis using the GeneChip Human Genome U133 Plus 2.0

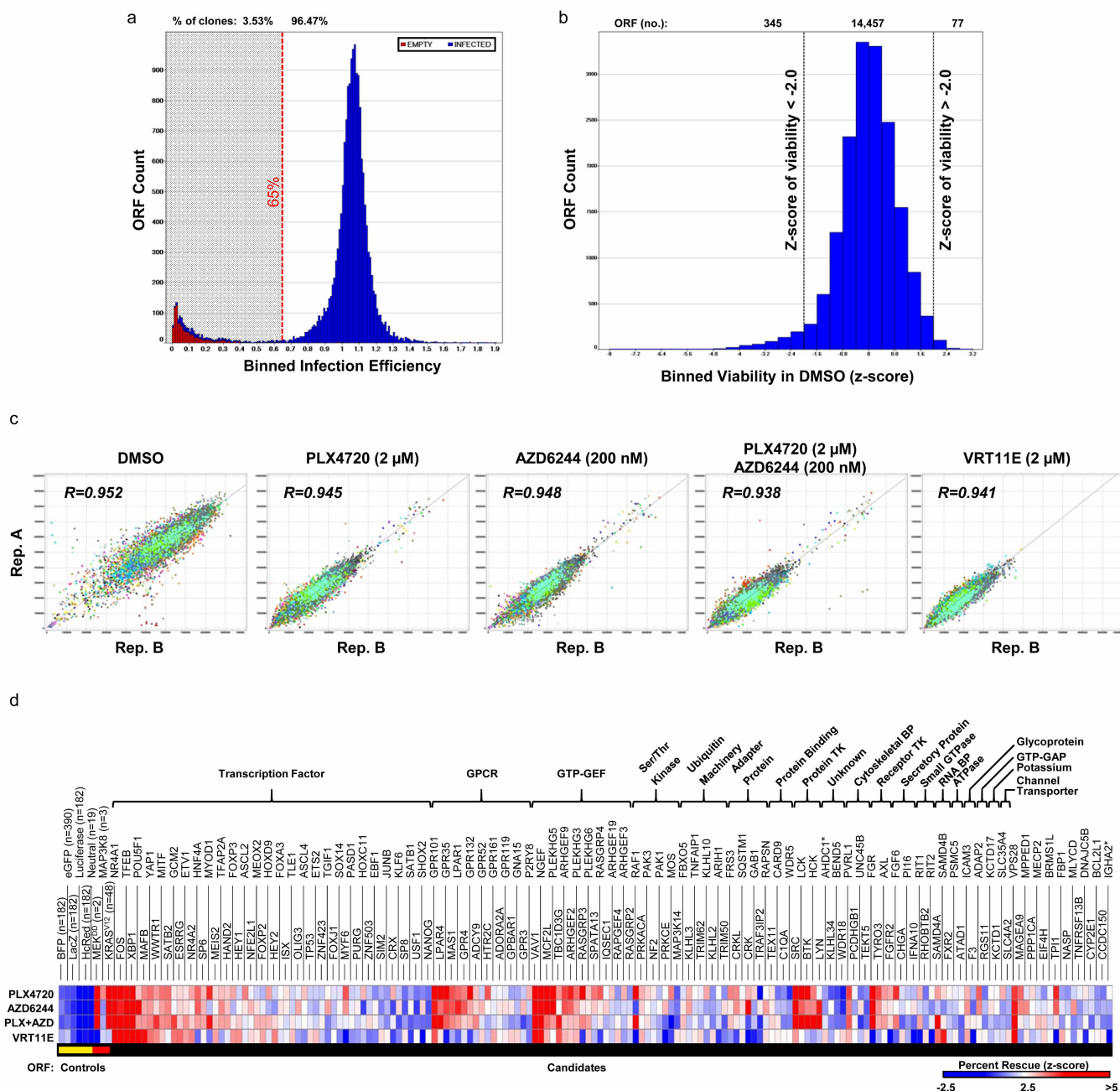
Affymetrix expression array (Affymetrix). Samples were converted to labelled, fragmented, cRNA per the Affymetrix protocol for use on the expression microarray. All expression arrays are available on the Broad-Novartis Cancer Cell Line Encyclopedia data portal at <http://www.broadinstitute.org/ccle/home> or on the Gene Expression Omnibus (GSE36133).

Melanoma tumour biopsies. Biopsied tumour material consisted of discarded and de-identified tissue that was obtained with informed consent and characterized under institutional review board (IRB) protocol 11-181 (Dana-Farber Cancer Institute). For paired specimens, 'on-treatment' samples were collected 10 to 14 days after initiation of MAPK inhibitor treatment (Extended Data Fig. 7e).



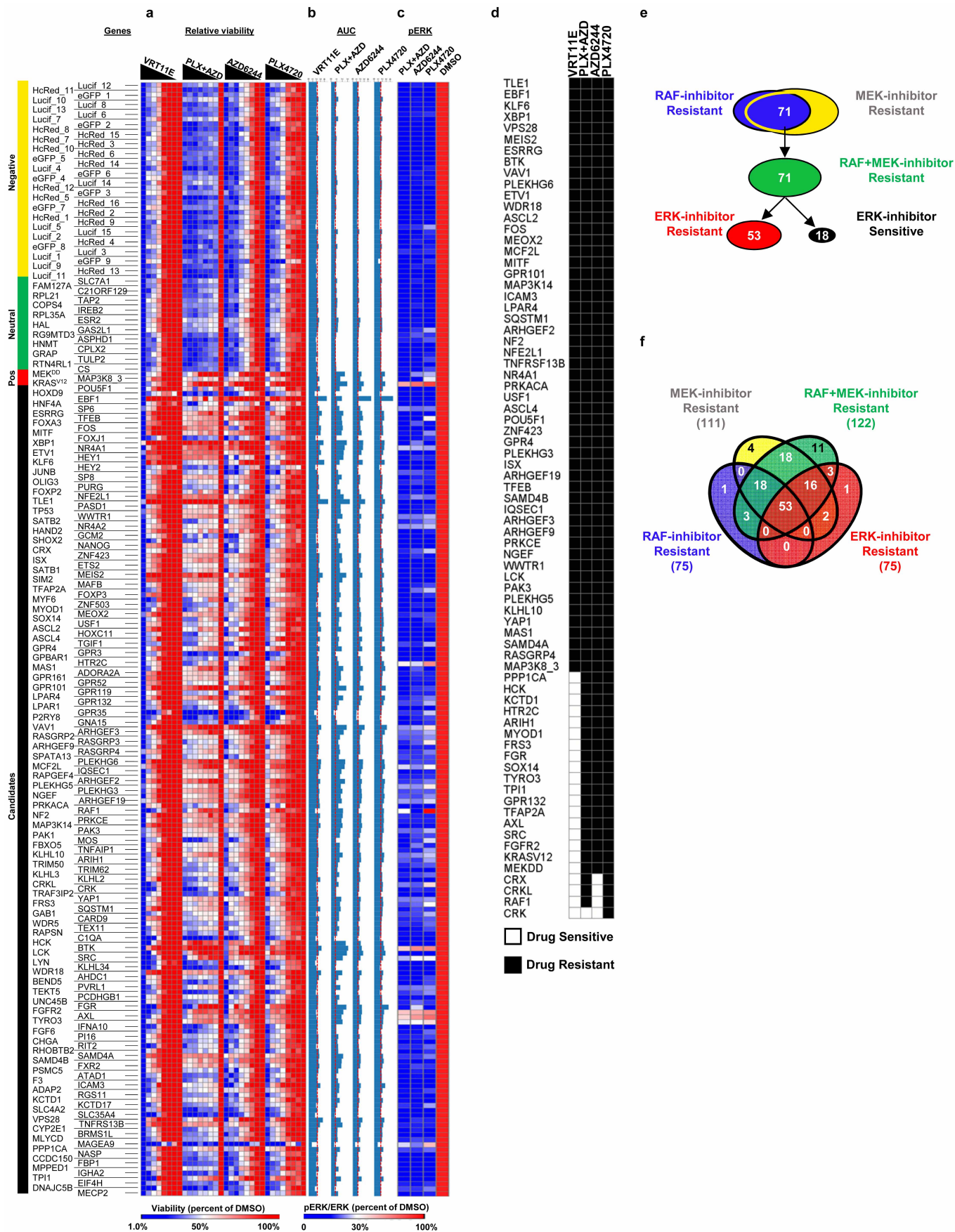
Extended Data Figure 1 | A systematic, functional approach to identifying drug-resistance genes. Schematic outlining the experimental approach taken to identify membrane-to-nucleus signalling pathways that mediate resistance

to MAPK-pathway inhibitors. Resulting data were used to identify gene networks capable of mediating drug resistance.



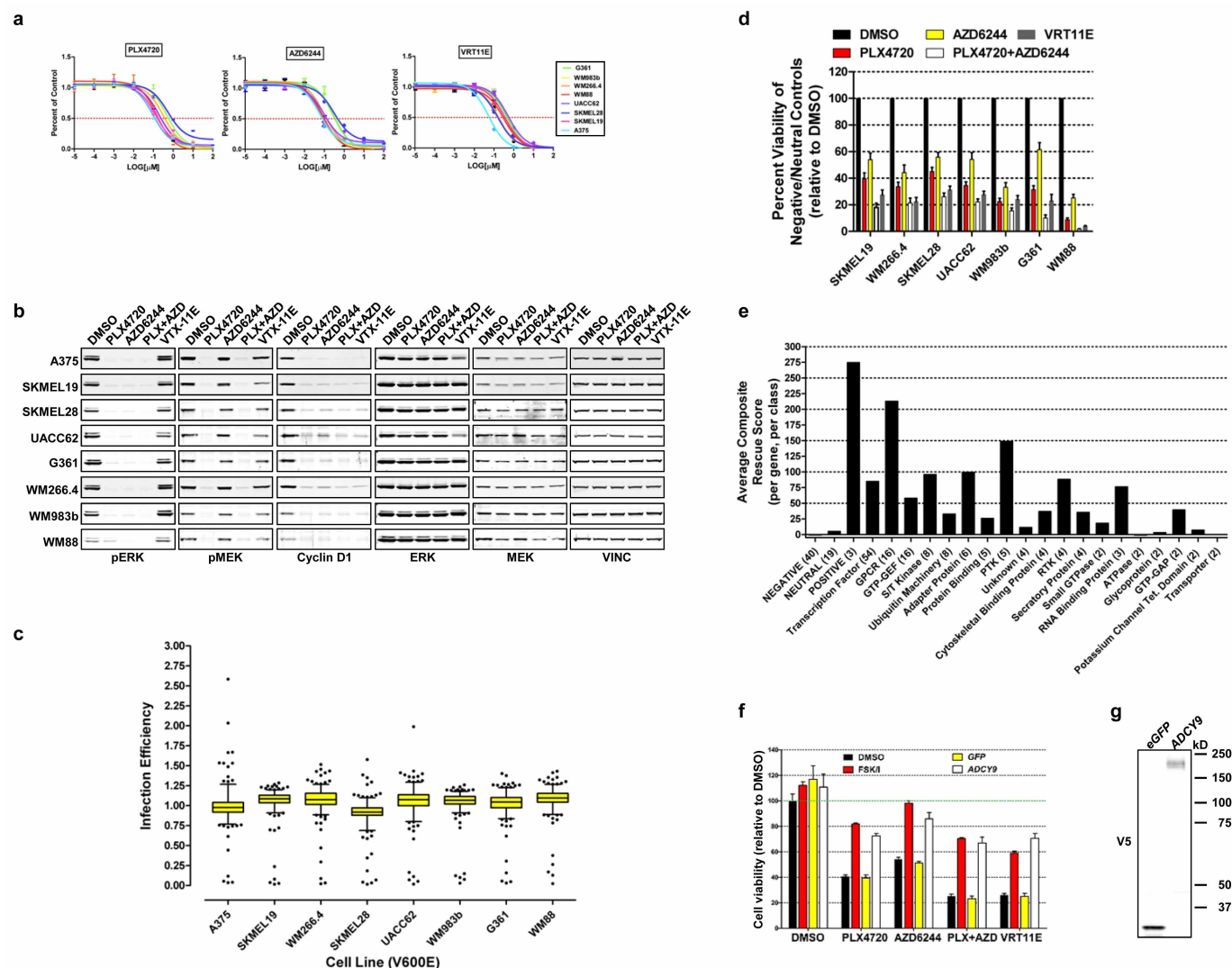
Extended Data Figure 2 | Near-genome-scale ORF and cDNA screens identify candidate MAPK-pathway inhibitor resistance genes. **a**, Histogram of infection efficiency in A375 cells observed in the primary resistance screens. Per cent of total ORFs above and below 65% infection efficiency are noted (red, dashed line). **b**, Histogram of the *z* score of A375-cell viability in DMSO observed in the primary resistance screen. Total ORFs above, below and within the indicated *z*-score thresholds are noted. **c**, Scatter plots and correlation (*R*) of

A375-cell viability (raw luminescence values) in the primary resistance screens. Colours distinguish viral screening plates. **d**, Heat map summary of controls and candidate resistance genes identified in primary resistance screens. Protein class and ORF class are indicated (positive control, red; negative control, yellow; experimental ORF, black). Asterisk identifies two genes whose empirical sequence is significantly divergent from its annotated reference sequence.



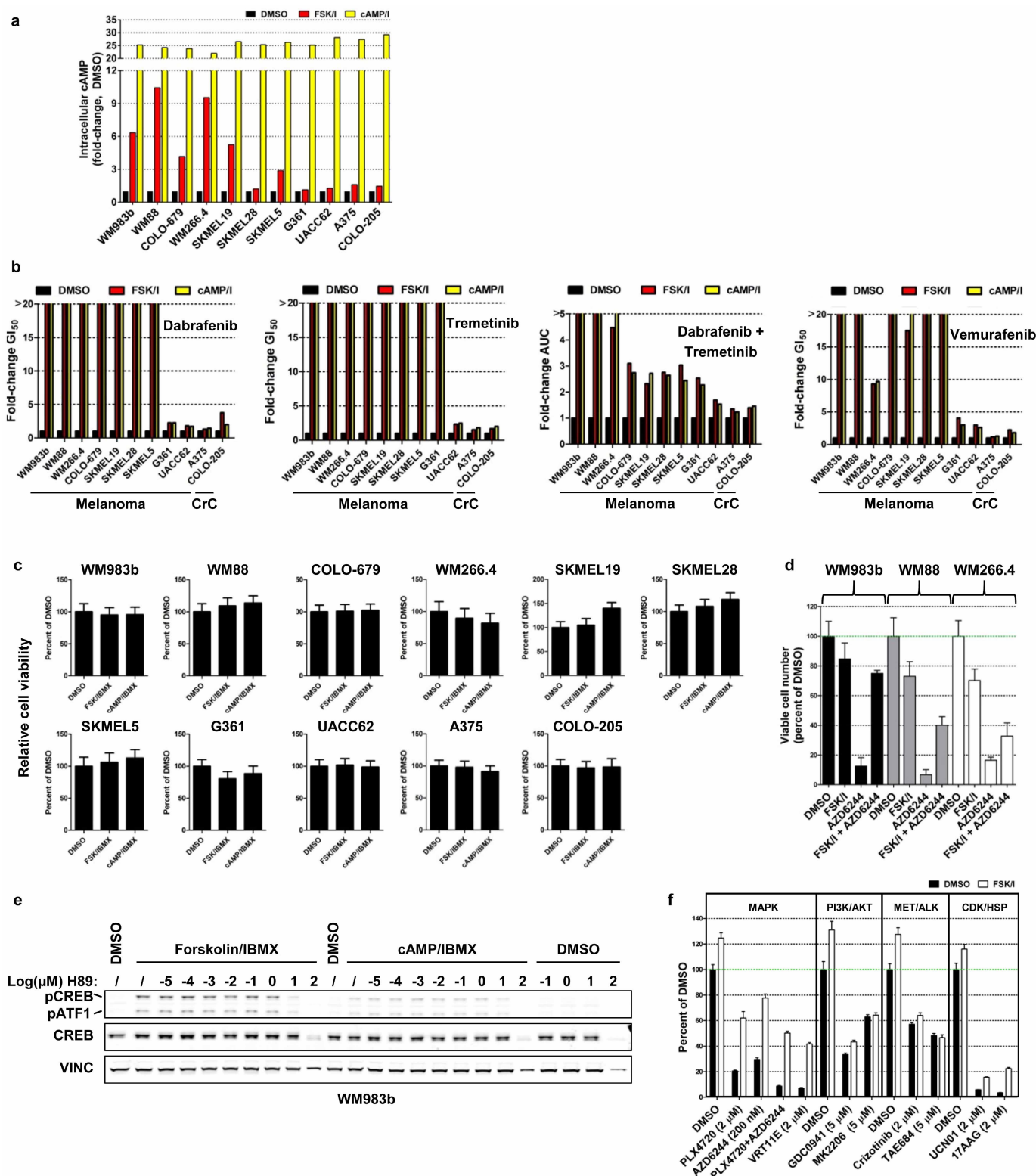
Extended Data Figure 3 | Patterns of drug resistance induced by candidate resistance genes. **a**, Heat map displaying the per cent rescue (viability in drug/viability in DMSO) for each candidate resistance ORF and control ORFs in the presence of log-fold concentrations of the indicated MAPK-pathway inhibitor. These data were used to generate drug-sensitivity curves. **b**, The area under the curve calculated for the drug sensitivity curves in **a** (red dashed lines denote significance thresholds). **c**, Heat map showing ERK phosphorylation data for all

candidate resistance genes and controls in A375 cells. **d**, Matrix of genes ectopically expressed in A375 cells (vertical axis) versus treatment condition (horizontal axis). Sensitivity is defined as yielding an area under the curve z score of <1.96 , resistance is defined as $z > 1.96$ ($P < 0.05$). **e**, Venn diagram showing the overlap of validated resistance genes, grouped by MAPK-pathway inhibitor, in A375 cells. **f**, Schematic showing the number of validated genes that confer resistance or sensitivity to indicated MAPK inhibitors.



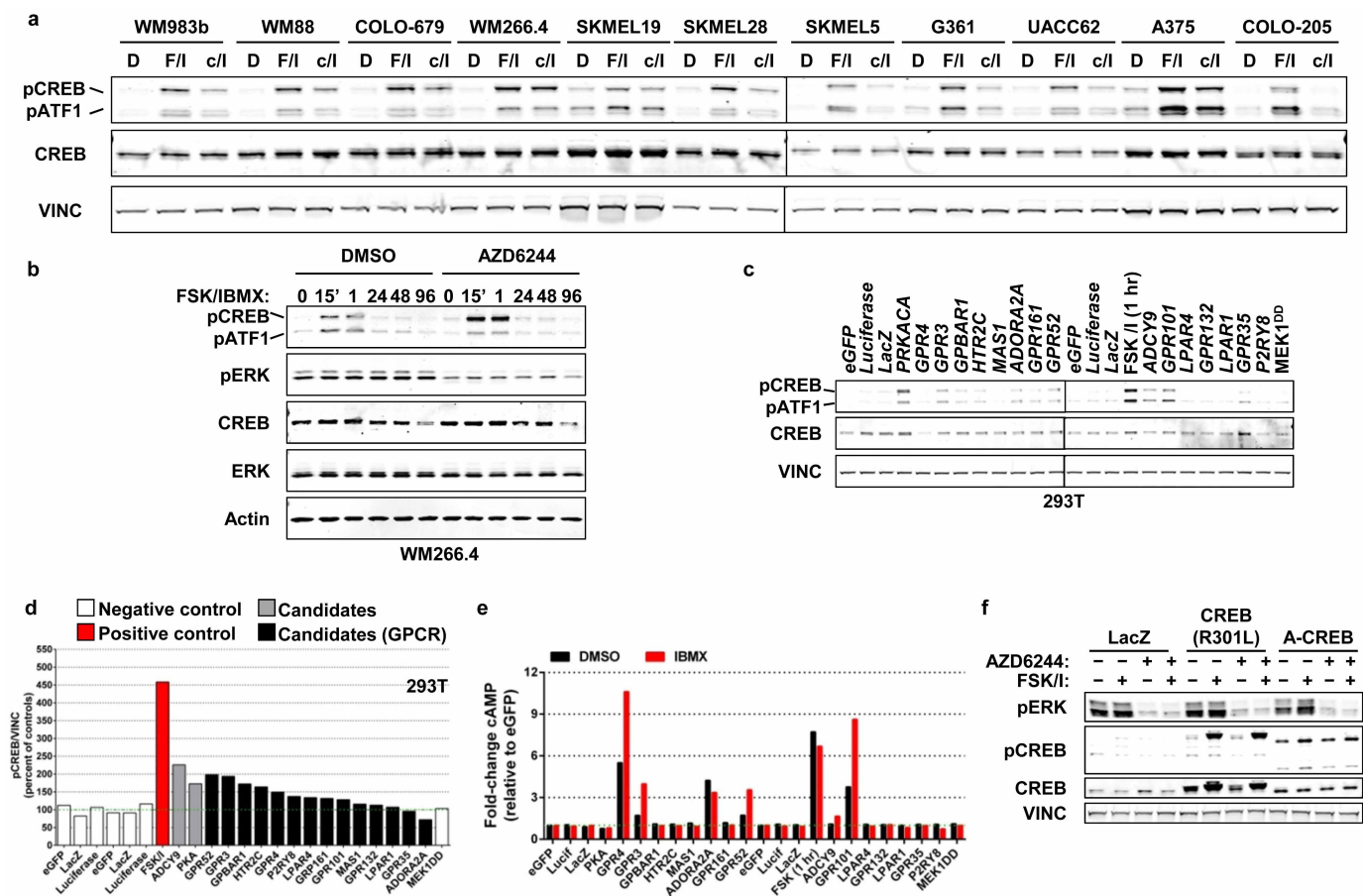
Extended Data Figure 4 | Broad validation of candidate resistance genes in a panel of BRAF(V600E)-mutant melanoma cell lines. **a**, Drug sensitivity curves for PLX4720, AZD6244 and VRT11E in the panel of eight BRAF(V600E)-mutant malignant melanoma cell lines used for the primary and validation screening experiments (described in Fig. 2). Error bars represent s.d. of mean, $n = 6$ technical replicates. **b**, Western blot analysis following treatment with indicated MAPK inhibitors in the panel of eight BRAF(V600E)-mutant malignant melanoma cell lines used in **a**. **c**, Box plot of all candidate and control ORF infection efficiencies in the panel of eight cell lines used in the validation screening experiments. Centre line represents the median value, box defines the 25th–75th percentile and whiskers define the 5–95% confidence interval. Outliers are shown as individual data points. **d**, Summary of the cellular viability (relative to DMSO) of negative and neutral control genes observed in validation screens. Bar graph shows the average viability (relative to

that of DMSO treatment) of each cell line when expressing the 59 negative and neutral control genes included in all validation screening experiments. Error bars represent s.d. of mean, each measured in technical duplicates. **e**, Average composite rescue score of each class of proteins identified among the resistance candidates (relates to Fig. 2). Number of genes within each protein class is shown in parentheses. **f**, *ADCY9* was identified as a resistance candidate in the primary resistance screen, but was a DNA failure in our independent prep of candidate virus. Therefore, *ADCY9* was not included in the high throughput validation screens, but was included in all subsequent validation work. These data show that *ADCY9* is able to confer resistance to all tested MAPK inhibitors to a similar degree as forskolin and IBMX treatment. Error bars represent s.d. of mean, $n = 6$ technical replicates. **g**, Western blot analysis of the expression of V5-epitope tagged eGFP and *ADCY9* in WM266.4.



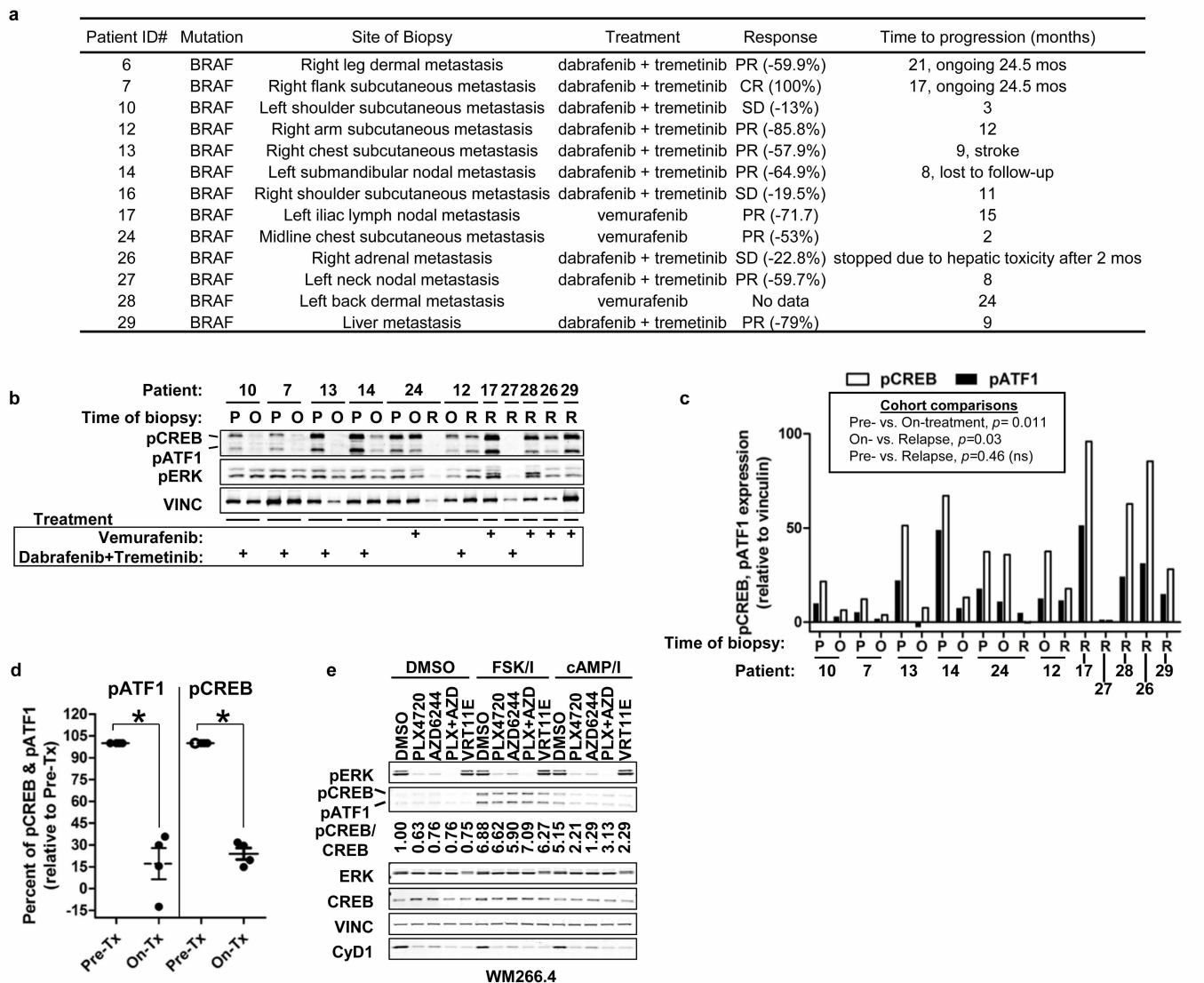
Extended Data Figure 5 | Cyclic AMP induces CREB and ATF1 phosphorylation and induces MAPK-pathway inhibitor resistance. **a**, Mean fold-change in intracellular cAMP following treatment with forskolin plus IBMX (FSK/I) or dibutyryl cAMP plus IBMX (cAMP/I) using a competitive cAMP ELISA ($n = 2$ technical replicates, representative of 2 independent experiments). **b**, Bar graphs showing the change in the half-maximal inhibitory concentration (GI_{50}) of BRAF(V600E)-mutant cell lines treated with escalating doses of indicated MAPK-pathway inhibitor in the presence of vehicle (DMSO), FSK/I or cAMP/I. **c**, Relative cell viability (per cent of DMSO)

following FSK/I or cAMP/I treatment in the absence of MAPK-pathway inhibitor treatment. Error bars represent s.d. of mean, $n = 8$ technical replicates. Data are representative of 2 independent experiments. **d**, Number of viable cells treated with the indicated compounds in the presence of vehicle (DMSO) or FSK/I. Error bars represent s.d. of mean, $n = 3$ technical replicates. **e**, Immunoblot analysis of WM983b cells following pre-treatment with the PKA inhibitor H89 and stimulation with FSK/I. **f**, Viability of WM266.4 cells treated with the indicated compounds and doses in the presence of vehicle (DMSO) or FSK/I. Error bars represent s.d. of mean, $n = 6$ technical replicates.



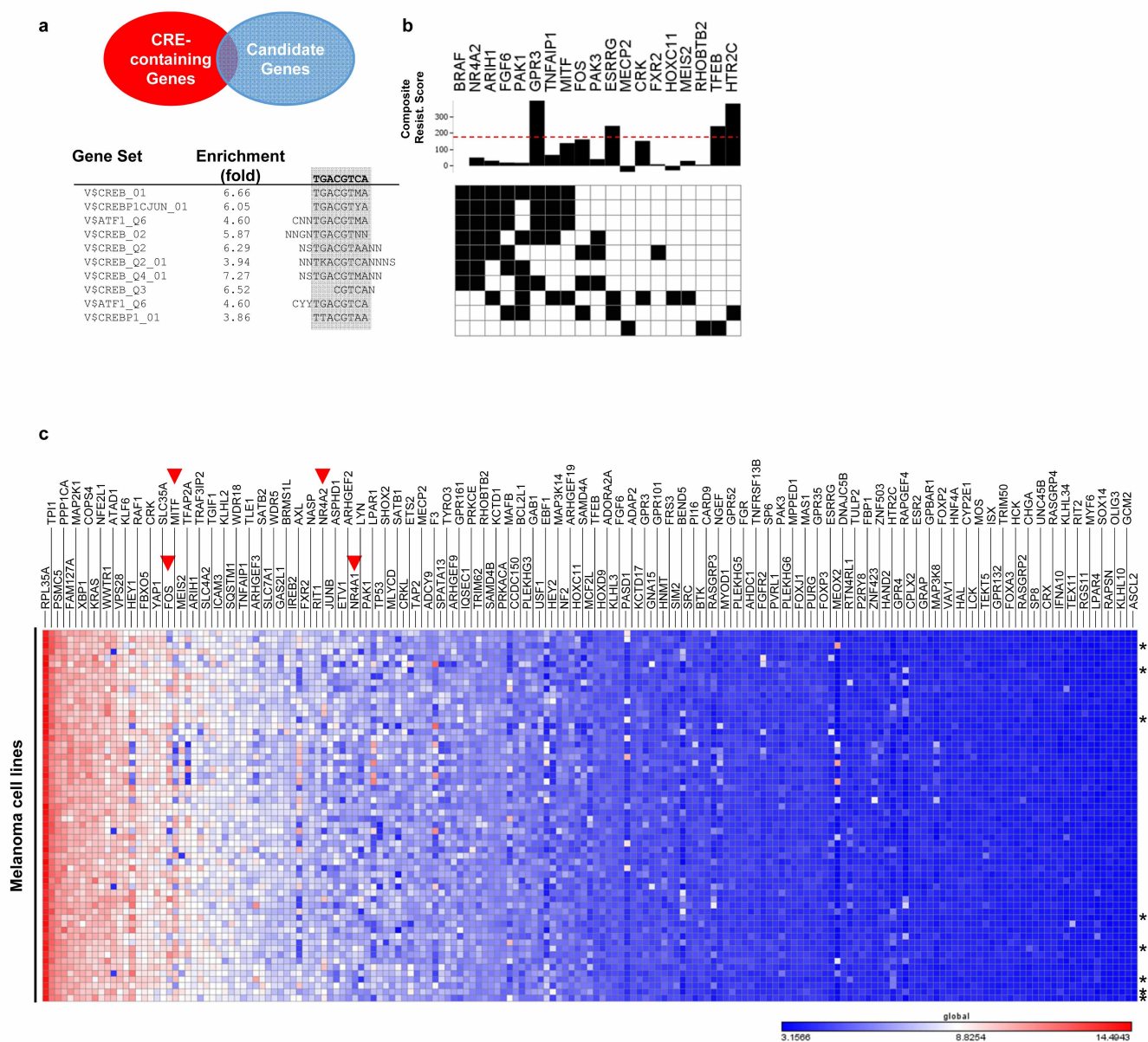
Extended Data Figure 6 | Candidate GPCR–PKA pathway genes induce cyclic AMP, and CREB and ATF1 phosphorylation. **a**, Western blot of BRAF(V600E)-mutant melanoma cell lines stimulated with forskolin and IBMX (FSK/I) or dibutyryl cAMP plus IBMX (cAMP/I). **b**, western blot analysis of WM266.4 cells treated with AZD6244, followed by stimulation with FSK/I. **c**, Western blot analysis of 293T lysates transfected with indicated genes or stimulated with FSK/I. **d**, Quantification of immunoblot analyses of 293T transiently transfected with the indicated expression constructs, pre-treated with IBMX (arbitrary units, $n = 2$ biological replicates). **e**, Mean control or

candidate gene-induced cAMP production was measured following transfection of 293T with indicated expression constructs or treatment with FSK/I. cAMP levels were determined using an immuno-competition assay in the presence (red bars) or absence (black bars) of IBMX ($n = 2$ technical replicates, data are representative of 3 independent experiments). The green dashed line represents levels of cAMP in negative controls (eGFP, luciferase, LacZ). **f**, Western blot analysis of WM266.4 cells expressing indicated constructs and treated with AZD6244 and/or FSK/I.



Extended Data Figure 7 | CREB activity is regulated in the context of drug treatment in patient biopsies. **a**, Summary of patient sample characteristics. **b**, Immunoblot analysis of lysates extracted from BRAF(V600E)-mutant human tumours biopsied pre-initiation of treatment (P), following 10 to 14 days of MAPK-inhibitor treatment (on-treatment, O) or following relapse (R). MAPK-inhibitor therapy is noted (vemurafenib, RAF inhibitor; dabrafenib, RAF inhibitor; tremetinib, MEK inhibitor). **c**, Comparison of quantified

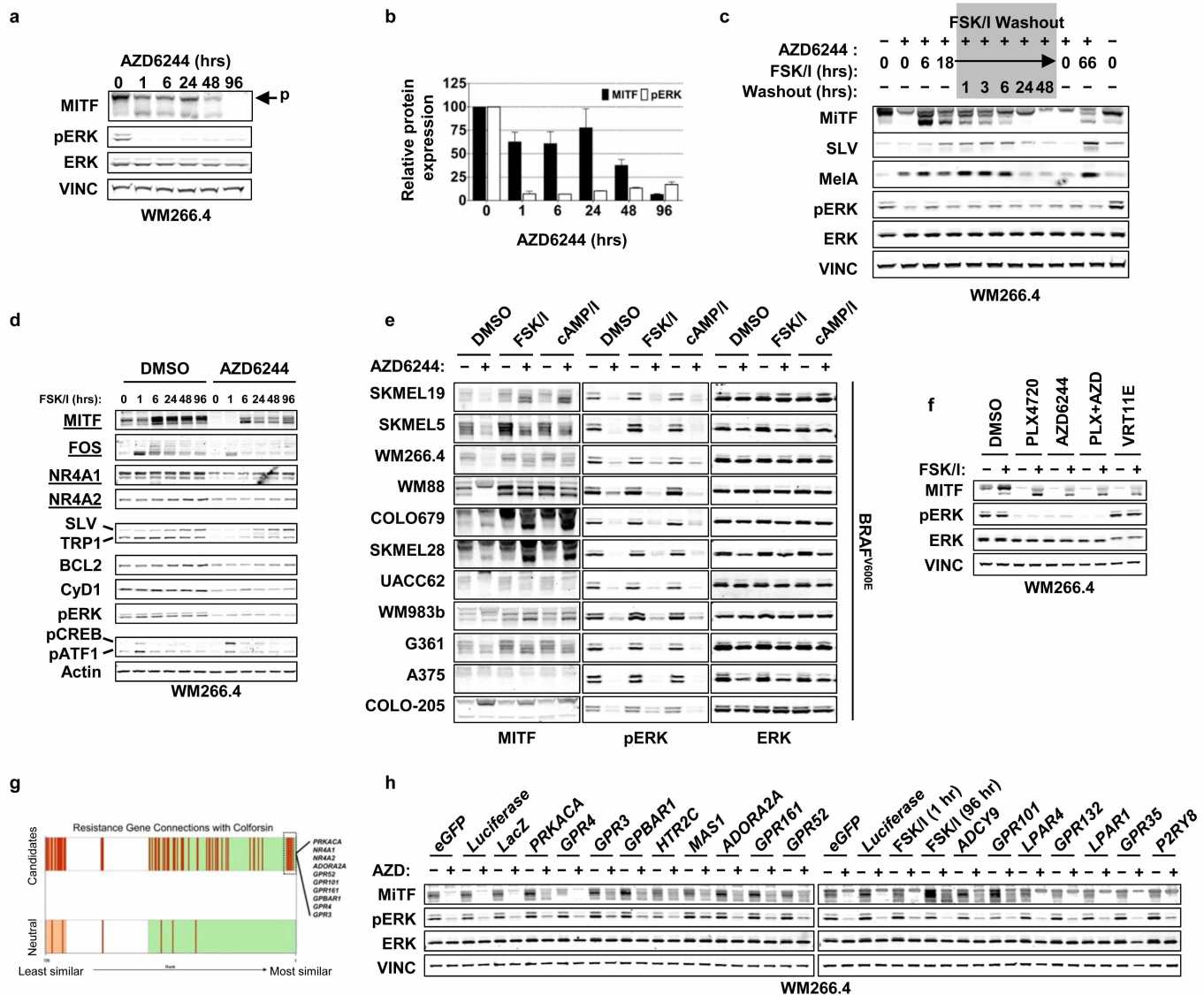
phosphorylated CREB (pCREB) and pATF1 from **b**, shown as individual tumours. **d**, Statistical analysis of pATF1 and pCREB as in **c**, normalized to pre-treatment levels. Samples analysed are restricted to the subset of the biopsies that are patient matched, lesion-matched and treatment-paired * $P < 0.0023$, by one-tailed t -test. **e**, Immunoblot analysis of WM266.4 cells following treatment with forskolin and IBMX (FSK/I) or dibutyryl cAMP and IBMX (cAMP/I) in the presence of vehicle (DMSO) or indicated MAPK inhibitors.



Extended Data Figure 8 | Identification of candidate resistance genes that are co-regulated by MAPK- and cAMP-PKA signalling pathways.

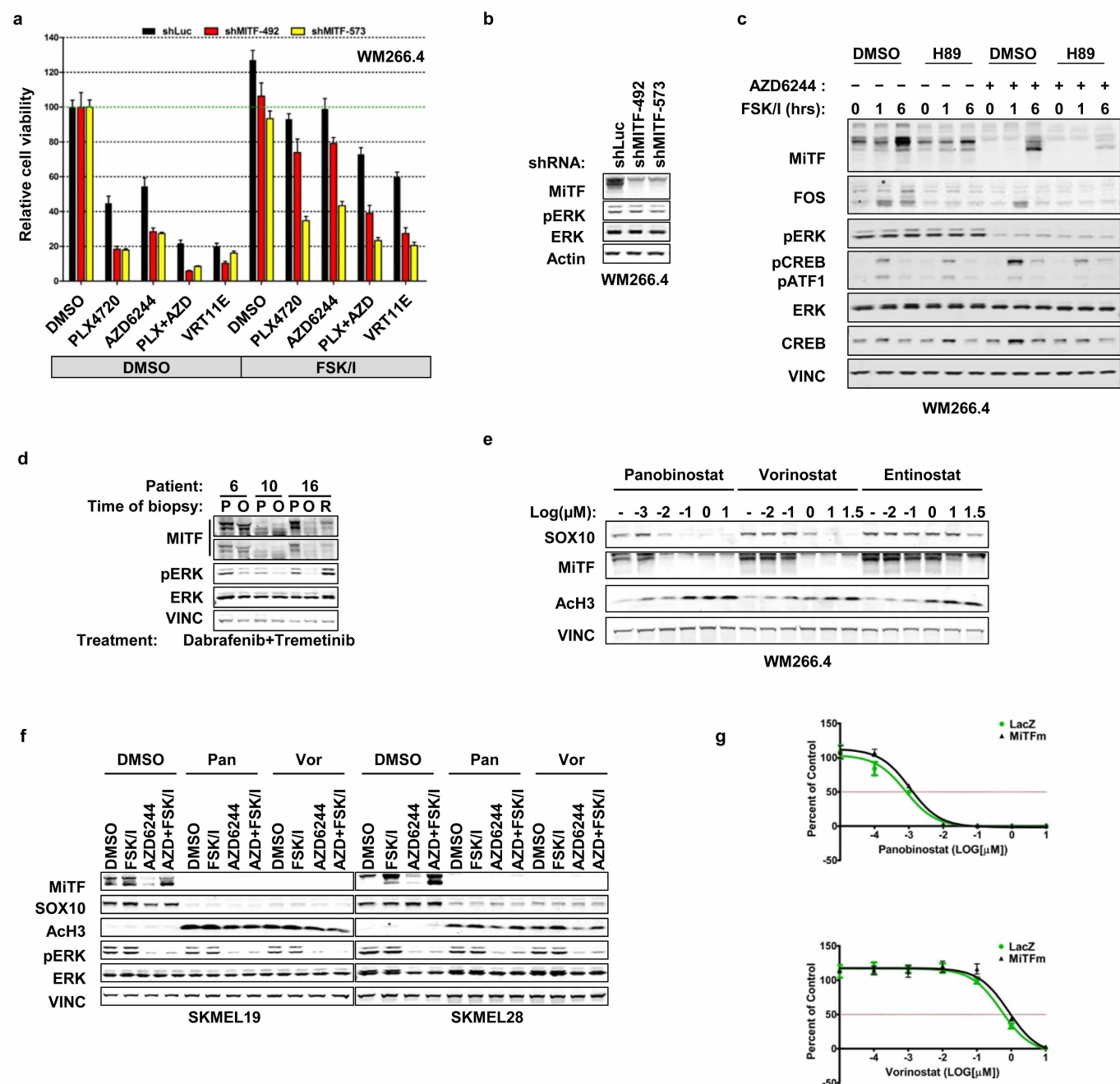
a, Candidate and neutral control genes containing cAMP response elements (CREs) were identified using gene sets extracted from MSigDB. Fold enrichment of the percentage of CRE-containing genes in candidates in relation to all genes screened for each gene set are noted. **b**, Matrix of CRE and candidate genes indicates the presence (black box) or absence (white box) of indicated

CRE. Composite resistance score for each gene (summarized in Fig. 2c) is noted. Red dashed line indicates a composite resistance score of 50. **c**, Global endogenous mRNA expression (Log₂ RMA) of candidate and neutral control genes across a panel of melanoma cell lines. Red arrows identify the four genes hypothesized to be regulated by both the MAPK-pathway and the cAMP-PKA-CREB pathway in melanoma: *MITF*, *FOS*, *NR4A1* and *NR4A2*. Asterisks identify the subset of cell lines used in for validation and primary screens.



Extended Data Figure 9 | cAMP–PKA regulation of MITF mediates resistance to MAPK pathway inhibition. **a**, Immunoblot analysis of WM266.4 cells treated as indicated. **b**, Quantification in lysates from WM266.4 cells treated as indicated. Arrow indicates the slower migrating, phosphorylated form of MITF. Error bars represent s.d. of mean, $n = 3$ biological replicates. **c**, Western blot analysis of WM266.4 cells following treatment with AZD6244 and stimulated for the indicated times with forskolin and IBMX (FSK/I). FSK/I was washed out of the cells and replenished with normal growth media. Cell lysates were collected at the indicated times. **d**, Immunoblot analysis of WM266.4 cells following treatment with FSK/I for the indicated times in the presence of vehicle (DMSO) or MEK inhibitors. Genes identified in resistance screens are underlined. **e**, Immunoblot analysis of a panel of BRAF(V600E)-

mutant malignant melanoma cell lines following treatment with AZD6244 in the presence of vehicle (DMSO), FSK/I or dibutyryl cAMP and IBMX (cAMP/I). **f**, Immunoblot analysis of WM266.4 cells following treatment with FSK/I in the presence of vehicle (DMSO) or indicated MAPK-pathway inhibitor. **g**, Gene signatures for all candidates and controls were generated in A375 cells and compared to the signatures of the cAMP-stimulating small molecule, colforsin. Individual genes are grouped as candidates or neutral controls, with each gene represented by a vertical line. Genes are ranked by similarity with colforsin, with number 1 being the most similar. A subset of the most similar genes is noted. **h**, Immunoblot analysis of WM266.4 cells after viral expression of the indicated genes or treatment with FSK/I in the presence of vehicle (DMSO) or AZD6244.



Extended Data Figure 10 | Inhibition of PKA or MITF impairs cAMP-mediated resistance to MAPK pathway inhibitors. **a**, Cell viability of WM266.4 cells expressing a control shRNA (shLuciferase) or shRNAs targeting MITF treated with indicated MAPK inhibitors and concomitant treatment with either DMSO or forskolin and IBMX (FSK/I). Error bars represent s.d. of mean, $n = 6$ technical replicates, data are representative of 2 independent experiments. **b**, Western blot analysis of WM266.4 cells expressing the shRNA constructs used in **a**. **c**, Western blot analysis of WM266.4 cells treated with AZD6244, followed by pre-treatment with DMSO or H89 and subsequent stimulation with FSK/I for the indicated times. **d**, Immunoblot analysis of lysates extracted from human BRAF(V600E)-positive melanoma biopsies.

Biopsies were obtained before treatment (P), on MAPK-inhibitor treatment for 10 to 14 days (on-treatment, O) or following relapse (R). **e**, Immunoblot analysis of WM266.4 cells treated with the indicated concentration of HDAC inhibitor. **f**, Immunoblot analysis of SKMEL19 and SKMEL28 in the presence of vehicle (DMSO) or AZD6244, followed by treatment with the indicated HDAC inhibitor (panobinostat; Pan, vorinostat; Vor) and subsequent stimulation with FSK/I. **g**, Drug-sensitivity curves of panobinostat and vorinostat in WM266.4 cells expressing LacZ or the melanocyte-specific isoform of MITF (MITFm). Error bars represent s.d. of mean, $n = 3$ technical replicates.

Oncogenic *Nras* has bimodal effects on stem cells that sustainably increase competitiveness

Qing Li¹, Natacha Bohin^{1*}, Tiffany Wen^{1*}, Victor Ng¹, Jeffrey Magee², Shann-Ching Chen³, Kevin Shannon⁴ & Sean J. Morrison²

'Pre-leukaemic' mutations are thought to promote clonal expansion of haematopoietic stem cells (HSCs) by increasing self-renewal and competitiveness¹; however, mutations that increase HSC proliferation tend to reduce competitiveness and self-renewal potential, raising the question of how a mutant HSC can sustainably outcompete wild-type HSCs. Activating mutations in *NRAS* are prevalent in human myeloproliferative neoplasms and leukaemia². Here we show that a single allele of oncogenic *Nras*^{G12D} increases HSC proliferation but also increases reconstituting and self-renewal potential upon serial transplantation in irradiated mice, all prior to leukaemia initiation. *Nras*^{G12D} also confers long-term self-renewal potential to multipotent progenitors. To explore the mechanism by which *Nras*^{G12D} promotes HSC proliferation and self-renewal, we assessed cell-cycle kinetics using H2B-GFP label retention and 5-bromo-deoxyuridine (BrdU) incorporation. *Nras*^{G12D} had a bimodal effect on HSCs, increasing the frequency with which some HSCs divide and reducing the frequency with which others divide. This mirrored bimodal effects on reconstituting potential, as rarely dividing *Nras*^{G12D} HSCs outcompeted wild-type HSCs, whereas frequently dividing *Nras*^{G12D} HSCs did not. *Nras*^{G12D} caused these effects by promoting STAT5 signalling, inducing different transcriptional responses in different subsets of HSCs. One signal can therefore increase HSC proliferation, competitiveness and self-renewal through bimodal effects on HSC gene expression, cycling and reconstituting potential.

To gain a durable competitive advantage, mutant HSCs must sustainably self-renew more frequently than wild-type HSCs. Yet increased HSC division is almost always associated with reduced self-renewal potential and HSC depletion^{3–5}. Many oncogenic mutations increase HSC proliferation but deplete HSCs, preventing clonal expansion⁶. Some oncogenic mutations do increase HSC self-renewal, including overexpression of *Ezh2* (ref. 7) or *Csf3r* truncation⁸, and deletion of *p18*^{INK4C} (ref. 9), *Tet2* (ref. 10), *Dnmt3a*¹¹ or *Lnk*^{12,13}. However, it remains uncertain whether these mutations can account for pre-leukaemic expansion.

Human leukaemias commonly have mutations that increase Ras signalling, including *NRAS* or *KRAS* point mutations². Mouse models with conditional expression of oncogenic *Kras*^{G12D} develop a rapid onset, aggressive myeloproliferative neoplasm (MPN)^{14,15}. *Kras*^{G12D} drives HSCs into cycle and reduces HSC frequency^{14,15}. *Nras*^{G12D} knock-in mice, on the other hand, develop an indolent MPN with delayed onset and prolonged survival^{16,17}. NF1 inactivation¹⁸ or *Nras*^{G12D} expression^{17,19} allow bone marrow cells to outcompete wild-type cells in transplantation assays, but it remains unclear whether they promote sustained pre-leukaemic expansion, or how that might occur.

To conditionally activate a single allele of *Nras*^{G12D} in HSCs we generated *Mx1-cre; Nras*^{G12D/+} mice in which the oncogenic *G12D* mutation was knocked into the endogenous *Nras* locus along with a floxed stop cassette²⁰. To induce *Nras*^{G12D} expression, mice were administered poly-inosine:poly-cytosine (pIpC) at 6–10 weeks after birth (Extended Data Fig. 1). At 2 weeks and 3 months after pIpC treatment,

more than twice as many *Nras*^{G12D/+} CD150⁺CD48[−]Lineage[−]Sca-1⁺c-kit⁺ (CD150⁺CD48[−]LSK) HSCs²¹ incorporated a 24-h pulse of BrdU as compared to control HSCs ($P < 0.01$; Fig. 1a). Consistent with this, twice as many *Nras*^{G12D/+} HSCs were in G1 phase of the cell cycle as compared to control HSCs (Extended Data Fig. 1b). This increase in HSC proliferation did not significantly affect the number of HSCs or multipotent progenitors (MPPs) 2 weeks after *Nras*^{G12D} activation (Fig. 1c). However, *Mx1-cre; Nras*^{G12D/+} mice had significantly more LSK cells in the bone marrow and spleen (Fig. 1c). We also observed a twofold increase in BrdU incorporation in HSCs, as well as an expansion of LSK cells in *Vav1-cre; Nras*^{G12D/+} mice as compared to controls (Fig. 1b; Extended Data Fig. 2a). Thus *Nras*^{G12D} increased HSC division and expanded the pool of primitive haematopoietic progenitors.

To test competitiveness we transplanted 5×10^5 whole bone marrow cells from *Mx1-cre; Nras*^{G12D/+} or control donors into irradiated wild-type recipients along with 5×10^5 recipient bone marrow cells. The *Nras*^{G12D/+} cells gave significantly higher levels of reconstitution than control cells in all lineages for at least 20 weeks after transplantation (Fig. 1d). In recipients of control donor cells, $69 \pm 13\%$ of HSCs, $47 \pm 12\%$ of MPPs and $44 \pm 12\%$ of LSK cells were donor-derived (Fig. 1e); however, in recipients of *Nras*^{G12D/+} donor cells, $93 \pm 8\%$ of HSCs, $90 \pm 8\%$ of MPPs, and $85 \pm 15\%$ of LSK cells were donor-derived (Fig. 1e). *Nras*^{G12D/+} HSCs therefore outcompeted wild-type HSCs.

To further test whether *Nras*^{G12D/+} HSCs could outcompete wild-type HSCs we transplanted 10^5 CD150⁺CD48[−]LSK donor HSCs from the bone marrow of *Mx1-cre; Nras*^{G12D/+} or littermate control mice (2 weeks after finishing pIpC) into irradiated wild-type recipients along with 3×10^5 recipient bone marrow cells. The *Nras*^{G12D/+} HSCs gave significantly higher levels of reconstitution compared to control donor HSCs in all lineages for at least 20 weeks after transplantation (Fig. 1f).

To assess self-renewal potential we serially transplanted 3×10^6 whole bone marrow cells from three or four recipients per treatment into 2 to 5 irradiated mice per recipient (depending on the number of bone marrow cells we recovered) during each round of transplantation. In secondary, tertiary and quaternary recipient mice we continued to observe significantly higher levels of reconstitution from the *Nras*^{G12D/+} donor cells than from control donor cells in all lineages (Fig. 2a–c). In tertiary recipient mice, the control cells gave only transient multilineage reconstitution as they appeared to exhaust their self-renewal potential. In contrast, the *Nras*^{G12D/+} HSCs gave high levels of long-term multilineage reconstitution in all 9 tertiary recipients, suggesting increased self-renewal potential. In quaternary recipient mice, *Nras*^{G12D/+} donor cells continued to give long-term multilineage reconstitution in most recipients whereas control donor cells gave only low levels of transient lymphoid reconstitution (Fig. 2c). *Nras*^{G12D} thus increased the self-renewal potential of HSCs in addition to increasing their rate of division (Fig. 1a) and their ability to compete with wild-type HSCs (Fig. 1d, f).

A fifth round of transplantation from four quaternary recipients of *Nras*^{G12D/+} cells did not yield any multilineage reconstitution by donor

¹Department of Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA. ²Howard Hughes Medical Institute, Department of Pediatrics, and Children's Research Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ³Department of Pathology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. ⁴Department of Pediatrics, University of California San Francisco, San Francisco, California 94158, USA.

*These authors contributed equally to this work.

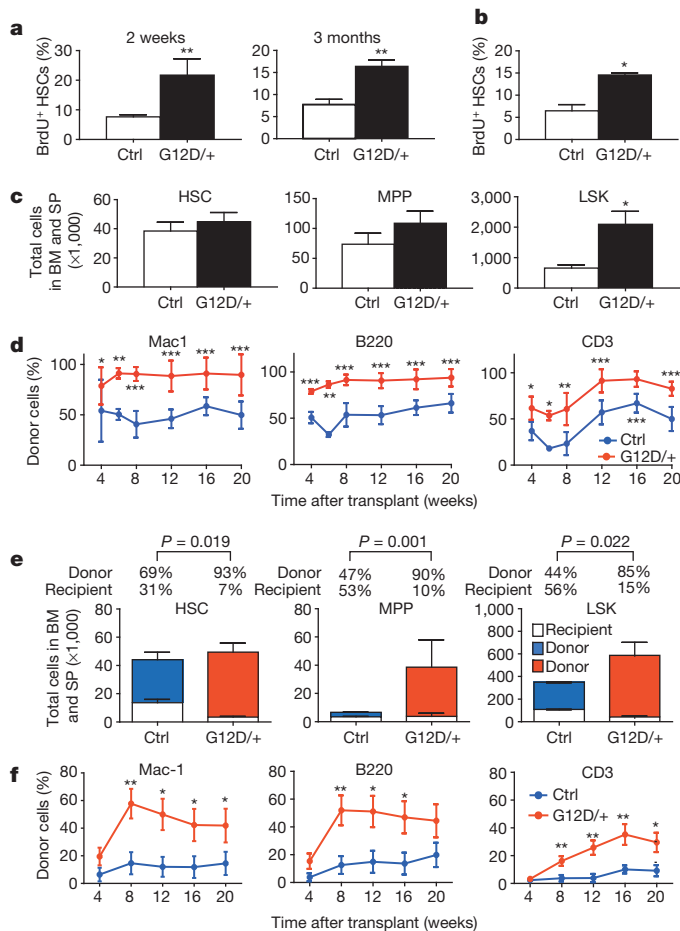


Figure 1 | *Nras*^{G12D/+} increased haematopoietic stem cell proliferation and competitiveness. **a**, A 24-h pulse of BrdU was administered to *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) and littermate control (Ctrl) mice at 2 weeks and 3 months after pIpC treatment (*n* = 3 mice per treatment). **b**, BrdU incorporation by CD150⁺CD48⁺ LSK HSCs from *Vav1-cre*; *Nras*^{G12D/+} mice (G12D/+) or littermate controls at 6–10-weeks of age (*n* = 3). **c**, The total number of CD150⁺CD48⁺ LSK HSCs, CD150⁺CD48⁺ LSK MPPs, and LSK cells in the bone marrow and spleens of *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) and littermate control mice at 2 weeks after pIpC treatment (*n* = 5 mice per treatment). **d**, Donor bone marrow cells (5×10^5) from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or littermate control mice at 2 weeks after pIpC treatment (*n* = 3 donors per genotype) were transplanted into irradiated recipient mice (*n* = 15 recipients per genotype) along with 5×10^5 recipient bone marrow cells. Donor cell reconstitution in the myeloid (Mac-1⁺ cells), B (B220⁺), and T (CD3⁺) cell lineages for 4 to 20 weeks after transplantation. **e**, Recipients of *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) bone marrow cells (*n* = 5) had significantly (*P* < 0.05) higher proportions of donor-derived HSCs, MPPs and LSK cells compared to recipients of control bone marrow cells. **f**, Ten donor HSCs from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or littermate control mice at 2 weeks after pIpC treatment (*n* = 3 donors per genotype) were transplanted into irradiated recipient mice (*n* = 14 recipients per genotype) along with 3×10^5 recipient bone marrow cells. Data represent mean \pm s.d. Two-tailed Student's *t*-tests were used to assess statistical significance. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

cells in the recipient mice (Extended Data Fig. 3). *Nras*^{G12D/+} HSCs therefore eventually exhausted their self-renewal potential despite self-renewing more than control cells. Serial transplantation of *Vav1-cre*; *Nras*^{G12D/+} whole bone marrow cells showed that *Nras*^{G12D/+} significantly increased HSC competitiveness in this genetic background as well (Extended Data Fig. 2b, c).

To test whether *Nras*^{G12D} expression influenced the reconstituting potential of MPPs, we transplanted 10 donor CD150⁺CD48⁺ LSK cells²² from the bone marrow of *Mx1-cre*; *Nras*^{G12D/+} or littermate control mice (2 weeks after finishing pIpC) into irradiated recipients along with

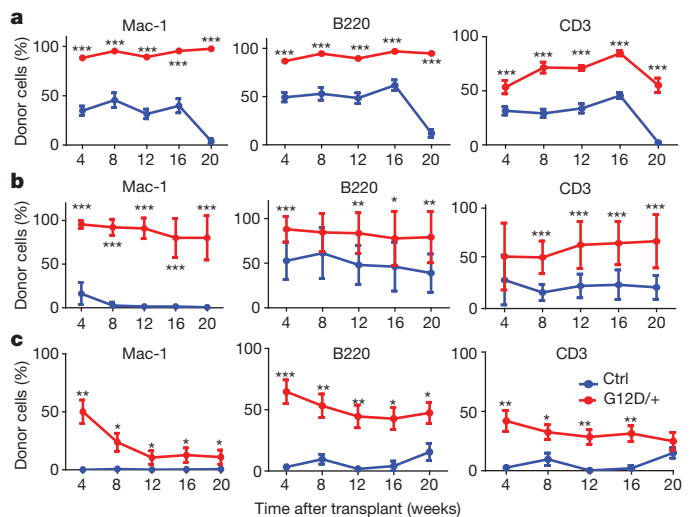


Figure 2 | *Nras*^{G12D/+} increased haematopoietic stem cell and multipotent progenitor self-renewal. **a**, Secondary transplantation (*n* = 19 recipients per genotype) of 3×10^6 bone marrow cells from primary recipient mice in Fig. 1c (*n* = 4 donors per genotype). Donor cell reconstitution in the myeloid (Mac-1⁺), B (B220⁺) and T (CD3⁺) cell lineages for 4 to 20 weeks after transplantation. **b**, Transplantation of 3×10^6 bone marrow cells from secondary recipient mice in Fig. 2a (*n* = 3 donors per genotype) into tertiary recipient mice (*n* = 8 recipients for control and 9 recipients for *Nras*^{G12D/+}). **c**, Transplantation of 3×10^6 bone marrow cells from tertiary recipient mice (*n* = 3 donors for control and 4 donors for *Nras*^{G12D/+}) in Fig. 2b into quaternary recipient mice (*n* = 7 recipients for control and 17 for *Nras*^{G12D/+}). Each serial transplant was performed at 20 weeks after the prior round of transplantation. Data represent mean \pm s.d. Two-tailed Student's *t*-tests were used to assess statistical significance. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

3×10^5 recipient bone marrow cells. Only one of 14 recipients of control MPPs exhibited long-term multilineage reconstitution by donor cells (Extended Data Fig. 4a). In contrast, 8 of 17 recipients of *Nras*^{G12D/+} MPPs were long-term multilineage reconstituted by donor cells. *Nras*^{G12D/+} MPPs were thus significantly (*P* < 0.01 across three independent experiments) more likely to give long-term multilineage reconstitution than control MPPs.

Nras^{G12D} did not detectably affect the reconstituting potential of 25 CD150⁺CD48⁺ LSK cells or 100 CD150⁺CD48⁺ LSK cells (which contain restricted myeloid progenitors²²) upon transplantation into irradiated mice (Extended Data Fig. 4b, c). *Nras*^{G12D/+} thus increases the self-renewal potentials of HSCs and MPPs but not necessarily other progenitors.

We did not detect any evidence of leukaemia or MPN in any of the recipient mice from the first, second, third or fourth rounds of serial transplantation in terms of blood cell counts (Extended Data Fig. 5) or histology (data not shown). Only two recipients of *Nras*^{G12D/+} cells and two recipients of control cells died spontaneously in these experiments. The effects of *Nras*^{G12D/+} on HSC function therefore occurred in the absence of leukaemogenesis.

To assess the effect of *Nras*^{G12D/+} on HSC cycling over time we mated the *Mx1-cre*; *Nras*^{G12D/+} mice with *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* double transgenic mice⁴. These mice allowed us to label HSCs with H2B-GFP during a 6-week period of doxycycline administration and then to follow the division history of all cells in the HSC pool as they diluted H2B-GFP with each round of division during a subsequent 12–15-week chase without doxycycline.

Two weeks after pIpC treatment, *Mx1-cre*; *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* mice and *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* controls (lacking *cre*) exhibited similar background levels of GFP fluorescence (Fig. 3a). After treating the mice with doxycycline for 6 weeks, all HSCs were strongly GFP⁺ (Fig. 3a). *Mx1-cre*; *Nras*^{G12D/+} and control HSCs exhibited indistinguishable levels of H2B-GFP labelling

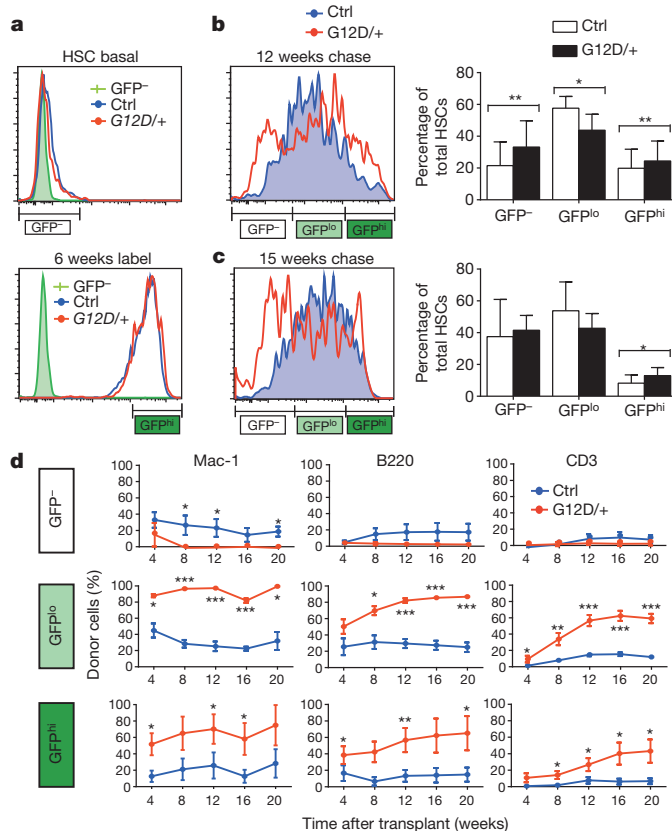


Figure 3 | *Nras*^{G12D/+} has a bimodal effect on HSC cycling.

a, GFP expression in HSCs from *Mx1-cre; Nras*^{G12D/+}; *Col1A1-H2B-GFP; Rosa26-M2-rtTA* mice (*G12D/+*) and littermate controls (Ctrl) without doxycycline treatment ($n = 3$, left), or after 6 weeks of doxycycline treatment ($n = 3$, right). **b**, GFP expression in HSCs from age- and sex-matched pairs of *Nras*^{G12D/+} and control mice after labelling followed by 12 weeks of chase without doxycycline ($n = 8$ pairs of mice from 8 independent experiments; $P < 0.05$ by two-way ANOVA and post hoc pairwise t -tests). Despite overlapping standard deviations, differences were statistically significant in pairwise t -tests because the frequencies of H2B-GFP⁺ HSCs and H2B-GFP^{hi} HSCs were always higher in the *Nras*^{G12D/+} mice. **c**, GFP expression in HSCs from pairs of age- and sex-matched *Nras*^{G12D/+} and control mice after 15 weeks of chase without doxycycline ($n = 7$ mice from 5 independent experiments). *Nras*^{G12D/+} mice always had higher frequencies of H2B-GFP^{hi} HSCs ($P < 0.05$ by pairwise t -tests). **d**, Transplantation of 15 CD150⁺CD48⁺LSK H2B-GFP^{hi} HSCs, 50 H2B-GFP^{lo} HSCs or 75 H2B-GFP⁺ HSCs from *Nras*^{G12D/+} or littermate control mice after 12 weeks of chase into irradiated wild-type recipients along with 3×10^5 recipient bone marrow cells (2 independent experiments with a total of 7 recipients per genotype). Data represent mean \pm s.d. Unless otherwise stated, two-tailed student's t -tests were used to assess statistical significance. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

immediately after doxycycline treatment (Fig. 3a). After a 12-week chase without doxycycline, the *Nras*^{G12D} and control HSCs exhibited a wide range of GFP expression levels (Fig. 3b). In contrast, most bone marrow cells from *Mx1-cre; Nras*^{G12D/+} and control mice exhibited GFP levels similar to background (Extended Data Fig. 6a). To assess the frequencies of HSCs that were most infrequently cycling, most frequently cycling, and moderately cycling we determined the frequencies of H2B-GFP^{hi} HSCs (with GFP levels similar to freshly labelled cells, see Fig. 3a), H2B-GFP⁺ HSCs (with little or no GFP expression above background, see Fig. 3a), and H2B-GFP^{lo} HSCs (with intermediate levels of GFP), respectively. In 8 independent experiments, *Nras*^{G12D} significantly ($P < 0.05$ by two-way ANOVA) increased the frequencies of both the H2B-GFP⁺ frequently cycling HSCs and the H2B-GFP^{hi} infrequently cycling HSCs in every pair of mice we examined ($n = 8$) (Fig. 3b). There was a corresponding significant decrease in the frequency of H2B-GFP^{lo} HSCs in *Nras*^{G12D} mice.

The median level of GFP fluorescence in H2B-GFP⁺ HSCs was significantly lower in *Nras*^{G12D/+} as compared to control mice (Extended Data Fig. 6b), suggesting that H2B-GFP⁺ *Nras*^{G12D/+} HSCs underwent more rounds of division on average. In contrast, the median level of GFP fluorescence in H2B-GFP^{hi} HSCs was significantly higher in *Nras*^{G12D/+} as compared to control mice (Extended Data Fig. 6b), suggesting that H2B-GFP^{hi} *Nras*^{G12D/+} HSCs tended to divide less than control H2B-GFP^{hi} HSCs on average. *Nras*^{G12D/+} thus had a bimodal effect, increasing the division of some HSCs and reducing the division of other HSCs.

We followed another cohort of age- and sex-matched pairs of *Mx1-cre; Nras*^{G12D/+} and control mice for 15 weeks after doxycycline removal. In five independent experiments, *Nras*^{G12D} significantly increased the frequency of H2B-GFP^{hi} HSCs in every pair of mice we examined ($n = 7$; $P < 0.05$) (Fig. 3c). We observed increased frequencies of H2B-GFP⁺ HSCs in the *Nras*^{G12D/+} mice from some pairs but not others, and overall the effect was not statistically significant (Fig. 3c). As the rapidly dividing subset of *Nras*^{G12D/+} HSCs differentiates more quickly than control HSCs (Fig. 3d), prolonged periods of chase after H2B-GFP labelling may not be appropriate to quantify the frequency of these cells. *Nras*^{G12D/+} significantly increased the rate at which MPPs divided (Extended Data Fig. 6c).

To test the relationship between division history and competitiveness we transplanted 15 CD150⁺CD48⁺LSK H2B-GFP^{hi} HSCs, 50 H2B-GFP^{lo} moderately cycling HSCs, or 75 H2B-GFP⁺ frequently cycling HSCs from *Nras*^{G12D/+} or control donor mice after 12 weeks of chase into irradiated wild-type recipients along with 3×10^5 recipient bone marrow cells. The *Nras*^{G12D/+} H2B-GFP⁺ frequently cycling HSCs gave significantly lower levels of donor cell reconstitution, at least in the myeloid lineages, as compared to control H2B-GFP⁺ HSCs (Fig. 3d). In contrast, the *Nras*^{G12D/+} H2B-GFP^{lo} and H2B-GFP^{hi} HSCs gave significantly ($P < 0.05$) higher levels of donor cell reconstitution in all lineages than the control H2B-GFP^{lo} and H2B-GFP^{hi} HSCs (Fig. 3d). *Nras*^{G12D/+} thus reduced the division and increased the competitiveness of some HSCs while increasing the division and reducing the competitiveness of other HSCs.

We continuously administered BrdU to *Mx1-cre; Nras*^{G12D/+} versus control mice beginning 2 weeks after pIpC treatment. We assessed the frequency of BrdU⁺ HSCs after 4, 10, 20 and 30 days of BrdU treatment (Extended Data Fig. 6d). Relative to control HSCs, significantly more *Mx1-cre; Nras*^{G12D/+} HSCs incorporated BrdU after 4 days ($32 \pm 0.1\%$ versus $24 \pm 1.2\%$, $P < 0.01$) and 10 days ($64 \pm 5.9\%$ versus $45 \pm 3.6\%$, $P < 0.02$) of BrdU administration. In contrast, significantly fewer *Mx1-cre; Nras*^{G12D/+} HSCs incorporated BrdU after 20 days ($78 \pm 4.7\%$ versus $86 \pm 0.7\%$, $P < 0.05$) and 30 days ($86 \pm 3.4\%$ versus $92 \pm 4.8\%$, $P < 0.02$) of BrdU administration. These data are consistent with the H2B-GFP label retention data in demonstrating that some *Nras*^{G12D/+} HSCs divide more frequently while other *Nras*^{G12D/+} HSCs divide less frequently than control HSCs.

We detected the activation of the canonical Ras effector, ERK, in bone marrow cells from *Mx1-cre; Nras*^{G12D/+} and *Mx1-cre; Nras*^{G12D/G12D} mice but not in LSK stem/progenitor cells or Lineage⁺c-kit⁺Sca-1⁺ myeloid progenitors (Extended Data Fig. 8a). We treated *Mx1-cre; Nras*^{G12D/+} or control mice with the MEK inhibitors, PD0325901 (5 mg per kg per day) or AZD6244 (25 mg per kg per day), and assessed the effects on BrdU incorporation in CD150⁺CD48⁺LSK HSCs. After eight days of treatment, splenocytes from PD0325901-treated mice of both genotypes showed reduced pERK levels (Extended Data Fig. 8b), but this did not affect the increased rate of BrdU incorporation by *Nras*^{G12D/+} HSCs (Extended Data Fig. 8c). In contrast, when we performed the same experiments with AZD6244, pERK activation was completely blocked in bone marrow and spleen (Extended Data Fig. 8d) and the increased cycling of *Nras*^{G12D/+} HSCs was abolished (Extended Data Fig. 8e). These data suggest that the more stringent inhibition of pERK activation by AZD6244 blocks the effect of *Nras*^{G12D/+} on HSC cycling.

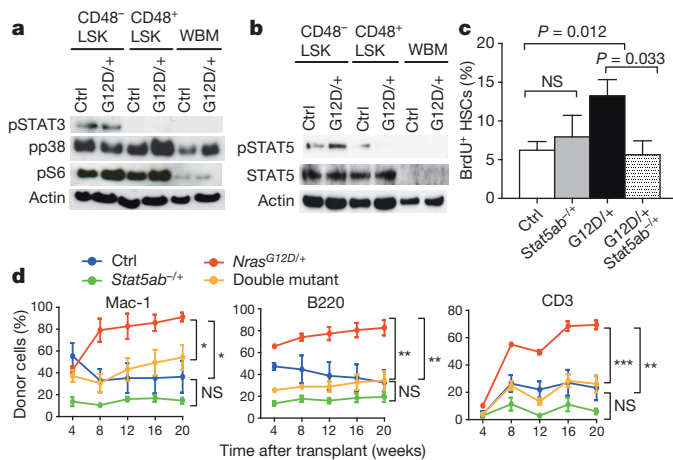


Figure 4 | Increased STAT5 activation mediates the effect of *Nras*^{G12D/+} on HSCs. **a**, **b**, Western blots for pSTAT3, pp38, pS6, and β -actin (**a**) and pSTAT5, total STAT5 and β -actin (**b**). Two additional experiments are shown in Extended Data Fig. 8j. Cells were stimulated in culture with stem cell factor and thrombopoietin for 30 min before protein extraction. **c**, The frequency of BrdU⁺ CD150⁺ CD48⁺ LSK HSCs after a 24-h pulse of BrdU to *Mx1-cre*; *Stat5ab*^{fl/+} (*Stat5ab*^{-/-}) mice, *Mx1-cre*; *Nras*^{G12D/+} (*G12D/+*) mice, *Mx1-cre*; *Nras*^{G12D/+}; *Stat5ab*^{fl/+} (*G12D/+*; *Stat5ab*^{-/-}) compound mutant mice, or control mice ($n = 4$). **d**, Donor bone marrow cells (5×10^5) from mice of each genotype were transplanted into irradiated recipients along with 5×10^5 recipient bone marrow cells (2 independent experiments with a total of 8 recipients per genotype). Data represent mean \pm s.d. Two-tailed student's *t*-tests were used to assess statistical significance.

We did not detect increased Akt (Extended Data Fig. 8f), S6, or p38 (Fig. 4a) phosphorylation in whole bone marrow cells, CD48⁺ LSK cells (HSCs and MPPs) or CD48⁺ LSK cells (mainly restricted progenitors²²) from *Mx1-cre*; *Nras*^{G12D/+} mice.

We performed gene expression profiling of H2B-GFP⁺ and H2B-GFP^{hi} CD150⁺ CD48⁺ LSK HSCs from 3 pairs of *Mx1-cre*; *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* mice and littermate controls 12 weeks after removal of doxycycline. Gene set enrichment analysis (GSEA) revealed that cell-cycle genes were significantly enriched in H2B-GFP⁺ *Nras*^{G12D/+} HSCs as compared to H2B-GFP⁺ control HSCs ($P < 0.0001$), but not in H2B-GFP^{hi} *Nras*^{G12D/+} HSCs as compared to H2B-GFP^{hi} control HSCs (Extended Data Fig. 7d). DNA replication genes and RNA polymerase genes were significantly enriched in H2B-GFP⁺ *Nras*^{G12D/+} HSCs as compared to H2B-GFP⁺ control HSCs ($P < 0.05$) but were significantly depleted in H2B-GFP^{hi} *Nras*^{G12D/+} HSCs as compared to H2B-GFP^{hi} control HSCs ($P < 0.01$; Extended Data Fig. 7e, f). These data demonstrate different transcriptional responses to *Nras* activation in H2B-GFP⁺ as compared to H2B-GFP^{hi} HSCs: genes associated with cell cycling were induced by *Nras* activation in the H2B-GFP⁺ HSCs but repressed by *Nras* activation in the H2B-GFP^{hi} HSCs.

Only two genes showed a similar change in expression with *Nras* activation in all of the multipotent cells we studied ($P \leq 0.05$ and fold change ≥ 2 ; Extended Data Fig. 7a–c). One of these, suppressor of cytokine signalling 2 (*Socs2*), was significantly reduced in expression in *Nras*^{G12D/+} as compared to control cells by microarray and quantitative PCR with reverse transcription (qRT-PCR) (Extended Data Figs 7c and 8g–i). *Socs2* negatively regulates STAT signalling in haematopoietic cells^{23,24}. We observed an increase in phosphorylated STAT5, but not STAT3, levels in *Nras*^{G12D/+} as compared to control CD48⁺ LSK cells, without affecting total STAT5 levels (Fig. 4a, b and Extended Data Fig. 8j).

To test whether the increased STAT5 phosphorylation in *Nras*^{G12D/+} CD48⁺ LSK cells increased HSC proliferation and self-renewal, we mated *Stat5ab*^{fl/+} mice²⁵ with *Mx1-cre*; *Nras*^{G12D/+} mice. Two weeks after finishing pIpC treatment, the western blot of flow cytometrically

sorted CD48⁺ LSK cells confirmed that deletion of one allele of *Stat5ab* in *Nras*^{G12D/+} HSCs reduced the levels of pSTAT5 and total STAT5 (Extended Data Fig. 8k). *Stat5ab*^{-/-} HSCs showed normal BrdU incorporation and *Nras*^{G12D/+} HSCs showed increased proliferation relative to control HSCs (Fig. 4c). Deletion of one allele of *Stat5ab* in *Nras*^{G12D/+} HSCs significantly reduced the rate of BrdU incorporation ($P < 0.05$). A reduction in STAT5 levels thus rescued the effects of *Nras*^{G12D} on HSC cycling. Neither deletion of *Stat5ab* nor activation of *Nras*^{G12D/+} significantly affected BrdU incorporation by the myeloid progenitors we examined (Extended Data Fig. 8l).

We next transplanted 5×10^5 whole bone marrow cells from control, *Mx1-cre*; *Nras*^{G12D/+}, *Mx1-cre*; *Stat5ab*^{-/-}, or *Mx1-cre*; *Nras*^{G12D/+}; *Stat5ab*^{-/-} (double mutant) donors (2 weeks after finishing pIpC treatment) into irradiated wild-type recipients along with 5×10^5 recipient bone marrow cells. *Nras*^{G12D/+} cells gave significantly higher levels of reconstitution than control cells in all lineages for at least 20 weeks after transplantation (Fig. 4d). Loss of one *Stat5ab* allele, reduced the level of reconstitution by donor cells relative to control cells but the difference was not statistically significant. In contrast, loss of a single allele of *Stat5ab* in the *Nras*^{G12D/+} background completely blocked the increased reconstitution by *Nras*^{G12D/+} cells such that levels of donor cell reconstitution were indistinguishable from control cells (Fig. 4d). An increase in STAT5 signalling is therefore required for increased competitiveness by *Nras*^{G12D} HSCs.

Nras^{G12D} is probably an early mutation in some leukaemias as it is widely observed in both MPN and myeloid leukaemias², and *Nras*^{G12D} mutations in mice lead only to a late onset MPN with prolonged survival^{16,17}. *NRAS* and *KRAS* mutations are frequently among the first mutations observed in pre-leukaemic clones that precede chronic myelomonocytic leukaemia (CMML)²⁶. Some juvenile myelomonocytic leukaemia (JMML) patients undergo remission, with or without therapy, yet continue to carry *NRAS* mutations in their haematopoietic cells^{27,28}. Germline *NRAS* mutations have also been reported in JMML patients²⁹, or patients with Noonan syndrome that develop JMML³⁰. The evidence that *NRAS* mutations can be found in normal haematopoietic cells, despite predisposing for the development of neoplasms, is consistent with our conclusion that they promote pre-leukaemic clonal expansion.

Our data provide a molecular explanation for how pre-leukaemic clonal expansion may occur. *Nras*^{G12D/+} has a bimodal effect on HSCs, increasing self-renewal potential and reducing division in one subset of HSCs while increasing division and reducing self-renewal in another subset of HSCs. Short-lived but rapidly dividing *Nras*^{G12D/+} HSCs presumably outcompete wild-type HSCs and are replenished over time by quiescent *Nras*^{G12D/+} HSCs that are slowly recruited into cycle. It will be interesting to determine whether the ability to induce bimodal responses in stem cell pools is a common feature of mutations that promote pre-malignant clonal expansion.

METHODS SUMMARY

Mice. All mice were housed in the Unit for Laboratory Animal Medicine at the University of Michigan and protocols were approved by the University of Michigan Committee on the Use and Care of Animals. *Nras*^{G12D/+} (ref. 20), *Stat5ab*^{fl/+} (ref. 25), *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA*⁺, *Vav1-cre*, and *Mx1-cre* mice were backcrossed for at least 10 generations onto a C57BL/Ka-CD45.2:Thy-1.1 background. Recipients in reconstitution assays were adult C57BL/Ka-CD45.1:Thy-1.2 mice, at least 8-weeks-old at the time of irradiation. pIpC (Amersham) was reconstituted in PBS and administered at 0.5 μ g per gram body mass per day by intraperitoneal injection. BrdU (Sigma) was administered as a single dose of 200 mg per kg body mass by intraperitoneal injection followed by 1 mg ml⁻¹ BrdU in the drinking water. For long term BrdU administration, BrdU water was changed every 3 days. Doxycycline (Research Products International) was added to the water at a concentration of 0.2% (m/v) along with 1% sucrose (Fisher). Both male and female mice were used in experiments and no randomization or blinding was performed. For all experiments, either littermates or age- and gender-matched mutants and control mice at 6–10-weeks of age were used.

Flow cytometry and HSC isolation. Bone marrow cells were harvested and CD150⁺ CD48⁺ Lin⁺ Sca1⁺ c-kit⁺ HSCs and CD150⁺ CD48⁺ Lin⁺ Sca1⁺ c-kit⁺ MPPs

were isolated as previously described^{21,22}. BrdU incorporation *in vivo* was measured by flow cytometry using the APC BrdU Flow Kit (BD Biosciences). To perform pyronin Y and DAPI staining, CD150⁺CD48⁺ LSK HSCs were sorted into 100% ethanol and placed in the cold room overnight. The cells were then washed with PBS and stained with pyronin Y (1 µg ml⁻¹) and DAPI (10 µg ml⁻¹) for 30 min before flow cytometric analysis.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 October 2012; accepted 7 November 2013.

Published online 27 November 2013.

- Rossi, D. J., Jamieson, C. H. & Weissman, I. L. Stem cells and the pathways to aging and cancer. *Cell* **132**, 681–696 (2008).
- Ward, A. F., Braun, B. S. & Shannon, K. M. Targeting oncogenic Ras signaling in hematologic malignancies. *Blood* **120**, 3397–3406 (2012).
- Essers, M. A. *et al.* IFN α activates dormant haematopoietic stem cells *in vivo*. *Nature* **458**, 904–908 (2009).
- Foudi, A. *et al.* Analysis of histone 2B–GFP retention reveals slowly cycling hematopoietic stem cells. *Nature Biotechnol.* **27**, 84–90 (2009).
- Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118–1129 (2008).
- Ross, L. *et al.* Less is more: unveiling the functional core of hematopoietic stem cells through knockout mice. *Cell Stem Cell* **11**, 302–317 (2012).
- Kamminga, L. M. *et al.* The Polycomb group gene *Ezh2* prevents hematopoietic stem cell exhaustion. *Blood* **107**, 2170–2179 (2006).
- Liu, F. *et al.* *Csf3r* mutations in mice confer a strong clonal HSC advantage via activation of Stat5. *J. Clin. Invest.* **118**, 946–955 (2008).
- Yuan, Y., Shen, H., Franklin, D. S., Scadden, D. T. & Cheng, T. *In vivo* self-renewing divisions of haematopoietic stem cells are increased in the absence of the early G1-phase inhibitor, p18^{INK4C}. *Nature Cell Biol.* **6**, 436–442 (2004).
- Moran-Crusio, K. *et al.* *Tet2* loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
- Challen, G. A. *et al.* *Dnmt3a* is essential for hematopoietic stem cell differentiation. *Nature Genet.* **44**, 23–31 (2012).
- Takizawa, H. *et al.* Enhanced engraftment of hematopoietic stem/progenitor cells by the transient inhibition of an adaptor protein, Lnk. *Blood* **107**, 2968–2975 (2006).
- Buza-Vidas, N. *et al.* Cytokines regulate postnatal hematopoietic stem cell expansion: opposing roles of thrombopoietin and LNK. *Genes Dev.* **20**, 2018–2023 (2006).
- Braun, B. S. *et al.* Somatic activation of oncogenic Kras in hematopoietic cells initiates a rapidly fatal myeloproliferative disorder. *Proc. Natl Acad. Sci. USA* **101**, 597–602 (2004).
- Sabnis, A. J. *et al.* Oncogenic Kras initiates leukemia in hematopoietic stem cells. *PLoS Biol.* **7**, e59 (2009).
- Li, Q. *et al.* Hematopoiesis and leukemogenesis in mice expressing oncogenic *Nras*^{G12D} from the endogenous locus. *Blood* **117**, 2022–2032 (2011).
- Wang, J. *et al.* Endogenous oncogenic *Nras* mutation promotes aberrant GM-CSF signaling in granulocytic/monocytic precursors in a murine model of chronic myelomonocytic leukemia. *Blood* **116**, 5991–6002 (2010).
- Zhang, Y., Taylor, B. R., Shannon, K. & Clapp, D. W. Quantitative effects of *Nf1* inactivation on *in vivo* hematopoiesis. *J. Clin. Invest.* **108**, 709–715 (2001).
- Wang, J. *et al.* *Nras*^{G12D/+} promotes leukemogenesis by aberrantly regulating hematopoietic stem cell functions. *Blood* **121**, 5203–5207 (2013).
- Haigis, K. M. *et al.* Differential effects of oncogenic K-Ras and N-Ras on proliferation, differentiation and tumor progression in the colon. *Nature Genet.* **40**, 600–608 (2008).
- Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
- Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102–116 (2013).
- Krebs, D. L. & Hilton, D. J. SOCS proteins: negative regulators of cytokine signaling. *Stem Cells* **19**, 378–387 (2001).
- Li, L. X., Goetz, C. A., Katerndahl, C. D., Sakaguchi, N. & Farrar, M. A. A. Flt3- and Ras-dependent pathway primes B cell development by inducing a state of IL-7 responsiveness. *J. Immunol.* **184**, 1728–1736 (2010).
- Cui, Y. *et al.* Inactivation of Stat5 in mouse mammary epithelium during pregnancy reveals distinct functions in cell proliferation, survival, and differentiation. *Mol. Cell Biol.* **24**, 8037–8047 (2004).
- Itzykson, R. *et al.* Clonal architecture of chronic myelomonocytic leukemias. *Blood* **121**, 2186–2198 (2013).
- Kotecha, N. *et al.* Single-cell profiling identifies aberrant STAT5 activation in myeloid malignancies with specific clinical and biologic correlates. *Cancer Cell* **14**, 335–343 (2008).
- Matsuda, K. *et al.* Spontaneous improvement of hematologic abnormalities in patients having juvenile myelomonocytic leukemia with specific RAS mutations. *Blood* **109**, 5477–5480 (2007).
- De Filippi, P. *et al.* Germ-line mutation of the *NRAS* gene may be responsible for the development of juvenile myelomonocytic leukaemia. *Br. J. Haematol.* **147**, 706–709 (2009).
- Kraoua, L. *et al.* Constitutional *NRAS* mutations are rare among patients with Noonan syndrome or juvenile myelomonocytic leukemia. *Am. J. Med. Genet. A* **158A**, 2407–2411 (2012).

Acknowledgements S.J.M. is a Howard Hughes Medical Institute Investigator and the Mary McDermott Cook Chair in Pediatric Genetics. This work was supported by the Cancer Prevention and Research Institute of Texas. Q.L. was supported by NIH K08-CA-134649 and V Foundation V Scholar award. Thanks to L. Hennighausen, K. Haigis and H. Hock for generously providing *Stat5ab*^{fl}, *Nras*^{G12D} and *Col1A1-H2B-GFP; Rosa26-M2-rtTA* mice. Thanks to M. Heeren and K. Rajan for help with genotyping and to R. Coolon and N. Vanderveen for mouse colony management.

Author Contributions Q.L. performed most of the experiments. N.B., T.W. and V.N. performed some of the experiments with help from Q.L. J.M. performed the western blot analysis of *Pten* mutant cells. S.C. performed statistical analysis of microarrays. Q.L., K.S., and S.J.M. conceived the project, designed experiments, interpreted results and wrote the manuscript.

Author Information Gene expression data have been deposited to the Gene Expression Omnibus with accession code number GSE45194. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Q.L. (lqing@umich.edu) or S.M. (Sean.Morrison@UTSouthwestern.edu).

METHODS

Mice. All mice were housed in the Unit for Laboratory Animal Medicine at the University of Michigan and protocols were approved by the University of Michigan Committee on the Use and Care of Animals. *Nras*^{G12D/+} (ref. 20), *Stat5ab*^{+/-} (ref. 25), *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* (ref. 4), *Vav1-cre* and *Mx1-cre* mice were backcrossed for at least 10 generations onto a C57BL/Ka-CD45.2:Thy-1.1 background. Recipients in reconstitution assays were adult C57BL/Ka-CD45.1:Thy-1.2 mice, at least 8-weeks-old at the time of irradiation. pIpC (Amersham) was reconstituted in PBS and administered at 0.5 µg per g of body mass per day by intraperitoneal (i.p.) injection. The 5-bromodeoxyuridine (BrdU, Sigma) was administered as a single dose of 200 mg per kg of body mass by i.p. injection followed by 1 mg ml⁻¹ BrdU in the drinking water. For long term BrdU administration, BrdU water was changed every 3 days. Doxycycline (Research Products International) was added to the water at a concentration of 0.2% (m/v) along with 1% sucrose (Fisher).

Statistical methods. Multiple independent experiments were performed to verify the reproducibility of all experimental findings. Group data always represents mean ± standard deviation. Unless otherwise indicated, two-tailed Student's *t*-tests were used to assess statistical significance. No randomization or blinding was used in any experiments. Experimental mice were not excluded in any experiments. In the case of measurements in which variation among experiments tends to be low (for example, HSC frequency) we generally examined between 3 to 6 mice. In the case of measurements in which variation among experiments tends to be higher (for example, reconstitution assays) we examined larger numbers of mice (7–20).

PCR of genomic DNA for genotyping. To assess the degree of *Nras*^{G12D} recombination in HSCs from *Mx1-cre*; *Nras*^{G12D/+} mice after pIpC treatment, bone marrow cells were harvested and stained for surface markers as described above. Single HSCs (CD150⁺CD48⁻ LSK cells) were sorted into 96-well plates containing methylcellulose medium (M3434, Stem Cell Technologies) and incubated for 14 days at 37 °C. Cells from each colony were resuspended in PBS then incubated with alkaline lysis buffer (25 mM NaOH, 0.2 mM EDTA), boiled, then neutralized by addition of an equal volume of neutralizing buffer (40 mM Tris-HCl). The neutralized extract was used for PCR with the following primers: F2, 5'-AGACGGGAGACTTGGCGAGC-3'; R1, 5'-GCTGGATCGTCAAGGCGCTTTCC-3'. To genotype mouse tail DNA for the presence of the *Nras*^{G12D} allele, primers R1 and F2 were used in addition to primer SD5', 5'-AGTAGCCACCATGGCTTGAGTAAGTCTGCA-3'. To genotype for the presence of the *Mx1-cre* transgene, primers F1 and R1 were used: F1 5'-ATTGCTGTCACTTGGTCGTGGC-3'; R1, 5'-GAAATGCTTCTGTCCGTTTG-3'. To check the presence of the *Rosa26-M2-rtTA* transgene, the following primers were used: 5'-AAAGTCGCTCTGAGTTGTAT-3'; 5'-GCGAAGAGTTTGTCTCAACC-3'; and 5'-GGAGCGGGA GAAATGGATATG-3'. To genotype mice for the presence of the *Col1A1-H2B-GFP* transgene, the following primers were used: 5'-CTGAAGTTCATCTGCAC CACC-3'; 5'-GAAGTTGTACTCCAGCTTGTGC-3'. To genotype mice for the deletion of *Stat5ab*, the following primers were used: 5'-GAAAGCAGCATGAAA GGGTTGGAG-3'; 5'-AGCAGCAACCAGAGGACTAC-3'; and 5'-AAGTTAT CTCGAGTTAGTCAGG-3'.

Flow cytometry and HSC isolation. Bone marrow cells were flushed from the long bones (tibiae and femurs) with Hank's buffered salt solution without calcium or magnesium, supplemented with 2% heat-inactivated calf serum (HBSS; Invitrogen). Cells were triturated and filtered through a nylon screen (70 µm; Sefar America) to obtain a single-cell suspension. CD150⁺CD48⁻ Lin⁻ Sca1⁺c-kit⁺ HSCs and CD150⁺CD48⁻ Lin⁻ Sca1⁺c-kit⁺ MPPs were isolated as previously described^{21,22}. For isolation of HSCs, whole bone marrow cells were incubated with antibodies to lineage (Lin) markers including B220 (6B2), CD3 (KT31.1), CD5 (53-7.3), CD8 (53-6.7), Gr-1 (8C5), CD41 (MWReg30) and Ter119 (Ter-119) that were conjugated to FITC, anti-CD150 antibody (TC15-12F12.2) conjugated to PE, anti-CD48 antibody (HM48-1) conjugated to PE-Cy7, anti-c-kit antibody (2B8) conjugated to APC, and anti-Sca1 antibody (D7) conjugated to PerCP/Cy5.5 (all antibodies were purchased from BioLegend). After washing, cells were incubated with anti-APC conjugated to paramagnetic microbeads (Miltenyi Biotec). The microbead bound (c-kit⁺) cells were then enriched using LS columns (Miltenyi Biotec). To identify CD45.2⁺ HSCs, antibodies against CD45.2 (104-FITC; BioLegend) and CD45.1 (A20-APC780, BioLegend) were used. Non-viable cells were excluded from sorts and analyses using the viability dye 4',6-diamidino-2-phenylindole (DAPI) (1 µg ml⁻¹). BrdU incorporation *in vivo* was measured by flow cytometry using the APC BrdU Flow Kit (BD Biosciences). To perform pyronin Y and DAPI staining, CD150⁺CD48⁻ LSK HSCs were sorted into 100% ethanol and placed at 4 °C overnight. The cells were then washed with PBS and stained with pyronin Y (1 µg ml⁻¹) and DAPI (10 µg ml⁻¹) for 30 min before flow cytometric analysis.

Long-term competitive repopulation assay. Adult recipient mice (CD45.1) were irradiated with an Orthovoltage X-ray source delivering approximately 300 rad min⁻¹ in two equal doses of 540 rad, delivered at least 2 h apart. Cells were injected into the retro-orbital venous sinus of anaesthetized recipients. Beginning 4 weeks after transplantation and continuing for at least 16 weeks, blood was obtained from the tail veins of recipient mice, subjected to ammonium-chloride potassium red cell lysis, and stained with directly conjugated antibodies to CD45.2 (104), CD45.1 (A20), B220 (6B2), Mac-1 (M1/70), CD3 (KT31.1) and Gr-1 (8C5) to monitor engraftment.

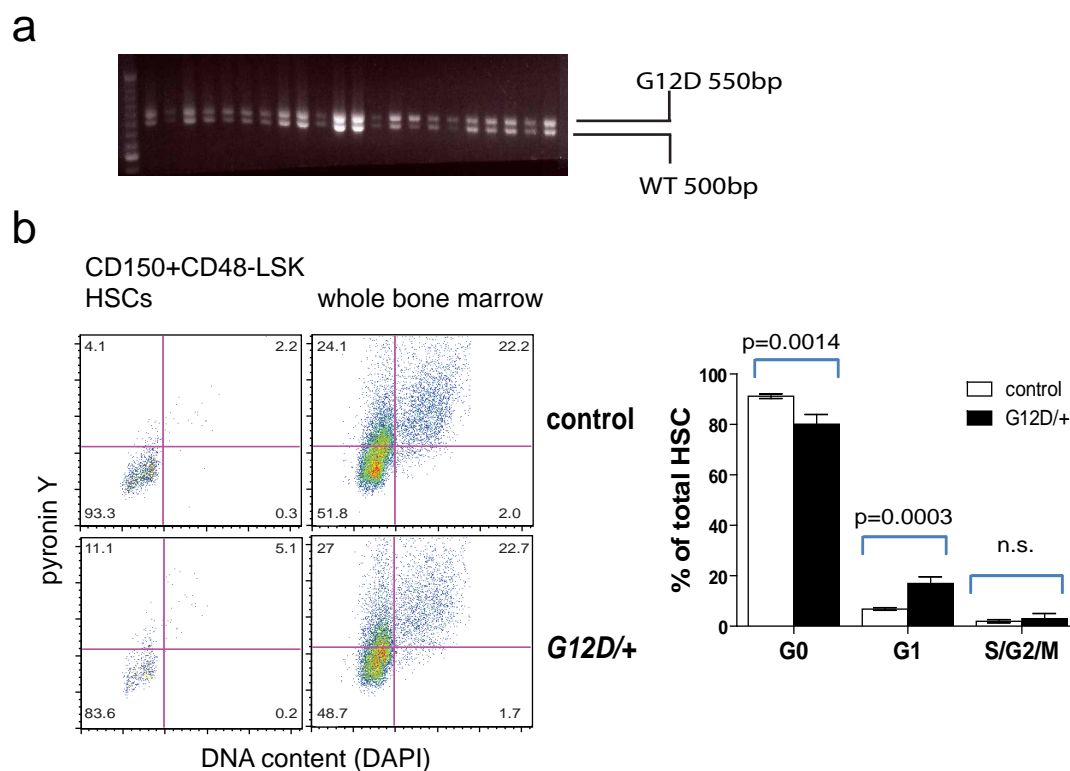
BrdU incorporation by myeloid progenitors. Two and a half hours after BrdU administration, whole bone marrow cells were incubated first with anti-c-kit antibody conjugated to biotin (2B8) then with antibodies to lineage (Lin) markers including B220 (6B2), CD3 (KT31.1), CD5 (53-7.3), CD8 (53-6.7), Gr-1 (8C5), CD41 (MWReg30) and Ter119 (Ter-119) that were conjugated to phycoerythrin (PE), anti-CD34 antibody (eBioscience, RAM34) conjugated to FITC, anti-CD16/CD32 antibody (93) conjugated to PE-Cy7, streptavidin conjugated to Alexa700 (Invitrogen S21383), and anti-Sca1 antibody (D7) conjugated to PerCP/Cy5.5 (all antibodies were purchased from BioLegend unless otherwise stated). BrdU incorporation was measured by flow cytometry using the APC BrdU Flow Kit (BD Biosciences).

Western blotting. The same number of cells (30,000 cells for CD48⁻ LSK cells or CD48⁺ LSK cells; 100,000 cells for LSK or Lineage⁻ c-kit⁺ Sca1⁻ cells) from each population to be analysed were sorted into HBSS with 2% FCS. The cells were then washed and incubated with 100 ng ml⁻¹ thrombopoietin and 100 ng ml⁻¹ stem cell factor at 37 °C for 30 min. The cells then were washed with PBS and precipitated with trichloroacetic acid (TCA) at a final concentration of 10% TCA. Extracts were incubated on ice for 15 min and spun down for 10 min at 16,100g at 4 °C. The supernatant was removed and the pellets were washed with acetone twice and then dried. The protein pellets were solubilized with solubilization buffer (9 M urea, 2% Triton X-100, 1% DTT) before adding lithium dodecyl sulphate loading buffer (Invitrogen). Proteins were separated on a Bis-Tris polyacrylamide gel (Invitrogen) and transferred to a polyvinylidene difluoride (PVDF) membrane (Millipore). All antibodies were purchased from Cell Signaling Technology. These include anti-pERK (T202/Y204), anti-pAkt (T308), anti-pS6 (S240/244), anti-pStat5 (Y694), anti-Stat3 (Y705), anti-ERK (137F5), anti-Stat5 (3H7), anti-pp38 (T180/Y182) and anti-β-actin (8H10D10).

Gene expression profiling. CD150⁺CD48⁻ LSK HSCs and CD150⁺CD48⁻ LSK MPPs were isolated by flow cytometry. Total RNA was isolated using Trizol (Invitrogen) followed by Qiagen RNeasy microkit purification according to the manufacturer's protocols. For microarray analysis, reverse transcription and linear amplification was performed on total RNA using the NuGen Ovation pico WTA system version 2 and then purified with the Qiaquick PCR purification kit (Qiagen). Six micrograms of amplified cDNAs were labelled with biotin using the Encore Biotin Module (NuGen) and submitted to the Microarray Core Facility of the University of Michigan Comprehensive Cancer Center for hybridization to Affymetrix Mouse Genome 430 2.0 Arrays. Statistical analyses were performed using R (ref. 31) version 2.15.2 and Bioconductor version 2.11 (ref. 32). Gene expression signals were normalized to the trimmed average of 500 using the Affymetrix MAS 5.0 algorithm. MAS5 signals less than 2 were set to 2 before log₂ transformation. Probe sets with MAS5 absent calls for all samples were excluded. Differential expression analysis was performed by limma³³ with estimation of fold change. Probe sets with limma *P*-value < 0.05 and fold change > 2 were considered differentially expressed. Gene Set Enrichment Analysis³⁴ was used to assess pathway enrichment. Gene expression data have been deposited to the Gene Expression Omnibus with accession number GSE45194 (<http://www.ncbi.nlm.nih.gov/geo/>).

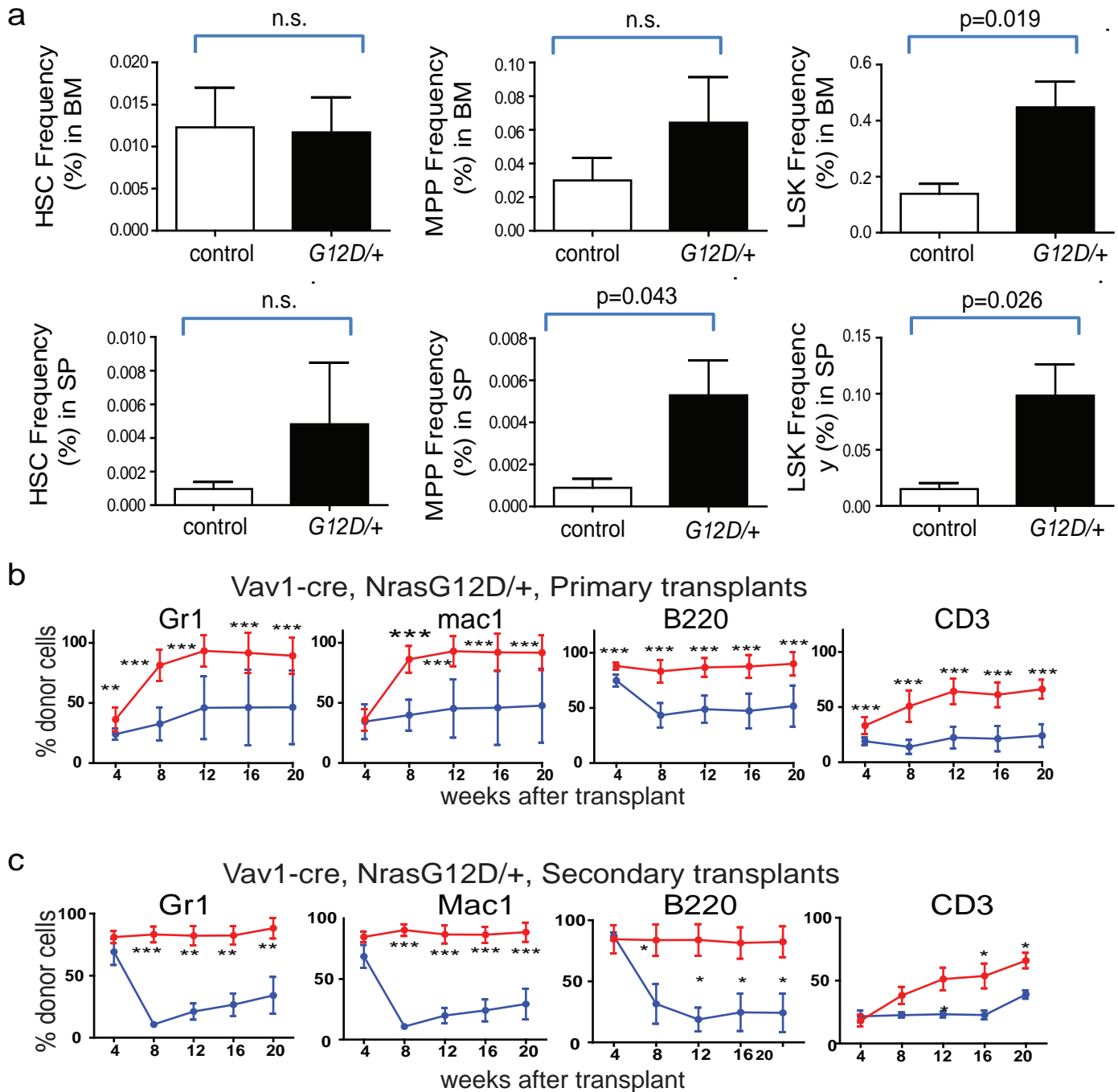
Quantitative RT-PCR. Total RNA was collected as described above and reverse transcription was performed with the High Capacity cDNA reverse transcription kit (Applied Biosystems). Real time PCR was performed with Absolute SYBR Green Rox mix (Thermo Scientific) using an ABI 7300 PCR machine. RNA from 100 cells was used for each reaction. Transcript levels were normalized to β-actin.

1. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org> (2009).
2. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
3. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
4. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).



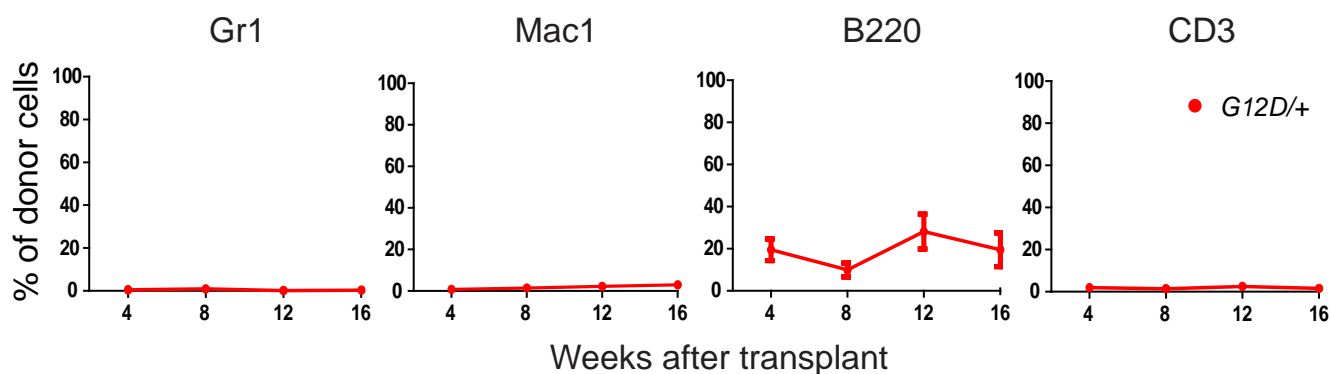
Extended Data Figure 1 | *Nras*^{G12D/+} increased HSC proliferation **a**, The *Nras*^{G12D} allele was recombined in all HSCs after 3 doses (every other day) of pIpC. Two weeks after the last dose of pIpC was administered to *Mx1-cre*; *Nras*^{G12D/+} mice, the mice were killed and individual CD150⁺CD48⁻LSK HSCs were sorted into methylcellulose cultures in 96-well plates. The cells were cultured for 14 days then DNA was extracted from individual colonies and genotyped by PCR. The size of the recombined *Nras*^{G12D} allele (G12D) was 550 base pairs (bp) and the *Nras*⁺ allele (wild-type, WT) was 500 bp. *Nras*

recombination was observed in 22 out of 22 HSC colonies examined. Blot is representative of three independent experiments. **b**, Cell cycle analysis of HSCs by pyronin Y and DAPI staining. CD150⁺CD48⁻LSK HSCs were sorted from *Mx1-cre*; *Nras*^{G12D/+} mice and littermate controls into 100% ethanol and stained with pyronin Y and DAPI to identify cells in G0 (left lower quadrant), G1 (left upper quadrant) and S/G2/M (right upper and lower quadrants). Data represent mean \pm s.d. Statistical analysis was performed with a two-way ANOVA ($P < 0.01$, $n = 4$) followed by pairwise post hoc *t*-tests.



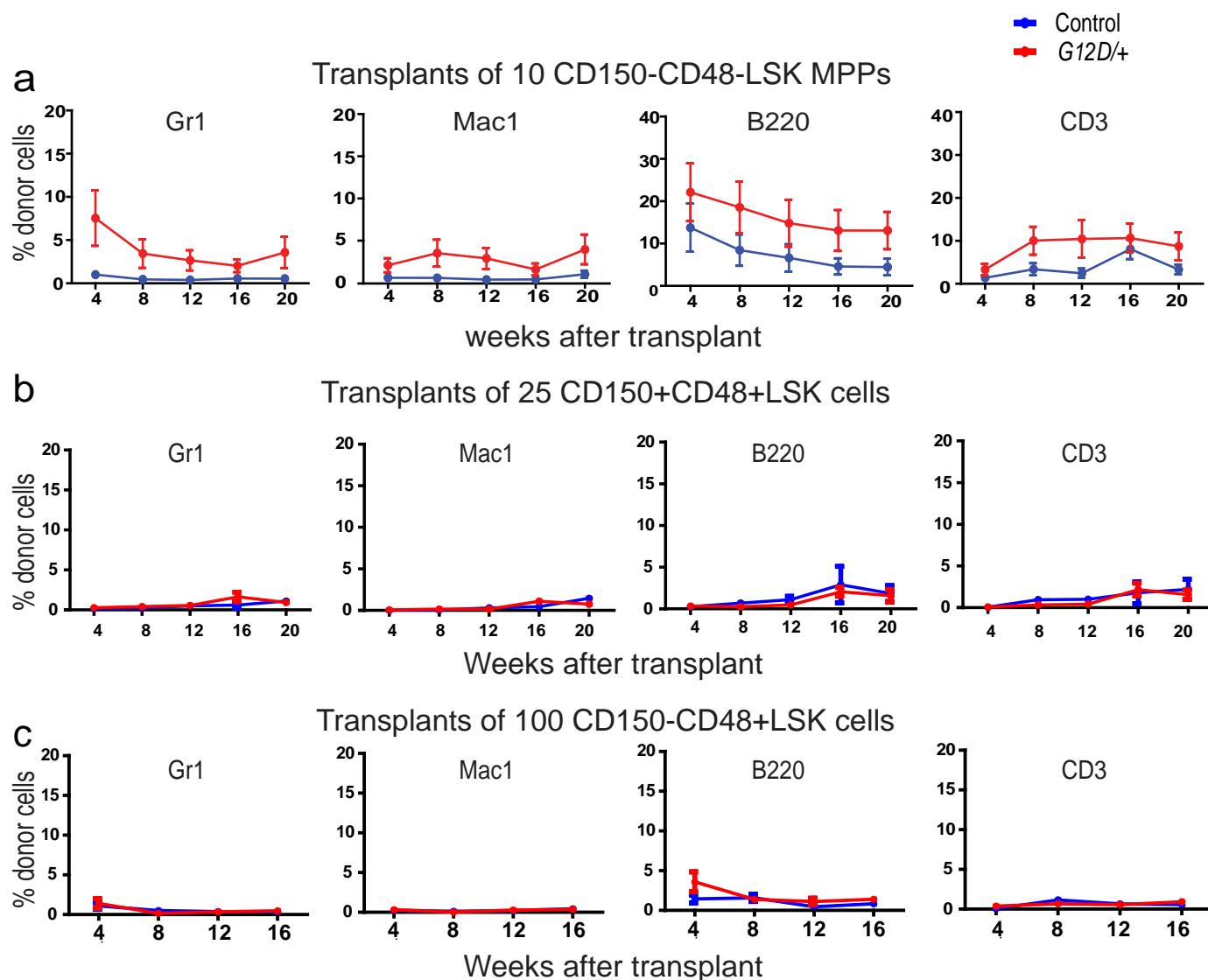
Extended Data Figure 2 | HSC competitiveness is increased in *Vav1*-cre; *Nras*^{G12D/+} mice. **a**, Frequencies of CD150⁺CD48⁺ LSK HSCs, CD150⁺CD48⁺ LSK MPPs, and LSK cells in the bone marrow (BM, top) and spleen (SP, bottom) of *Vav1*-cre; *Nras*^{G12D/+} (*G12D/+*) or littermate control mice (*n* = 4) at 6–10-weeks of age. **b**, Donor bone marrow cells (5×10^5) from *Vav1*-cre; *Nras*^{G12D/+} (*G12D/+*) or littermate control mice at 6–10-weeks of age were transplanted into irradiated recipient mice along with 5×10^5

recipient bone marrow cells (3 donors per genotype were each transplanted into 4 recipients per donor). **c**, Secondary transplantation of 3×10^6 bone marrow cells from primary recipient mice in Extended Data Fig. 2b at 20 weeks after transplantation (2 primary recipients per genotype were each transplanted into 4 secondary recipients per primary recipient). Data represent mean \pm s.d. Two-tailed Student's *t*-tests were used to assess statistical significance. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.



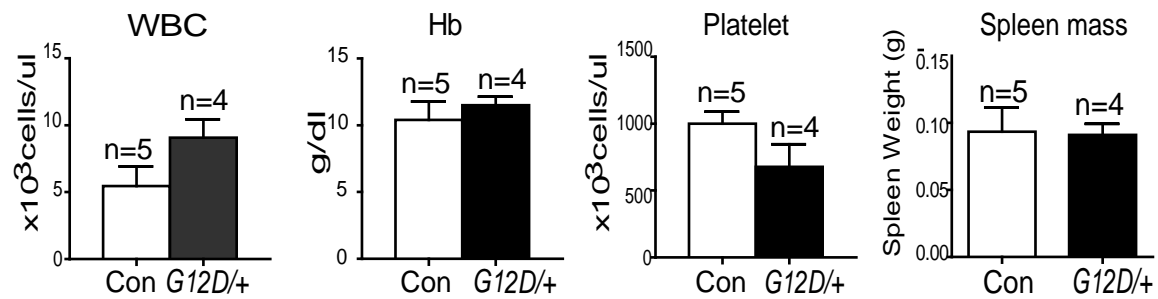
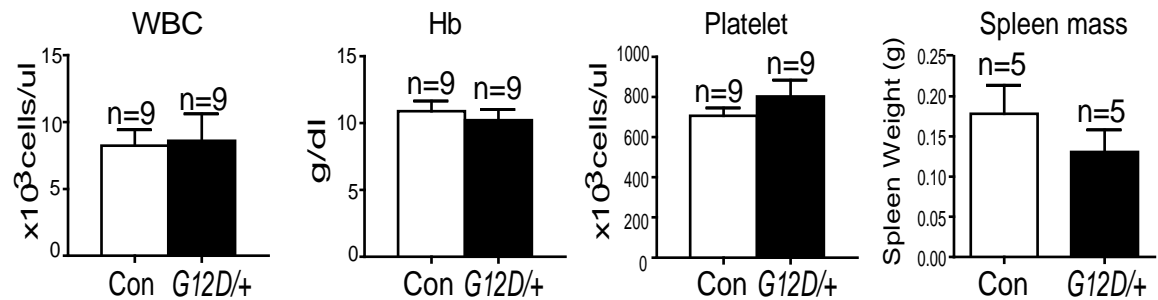
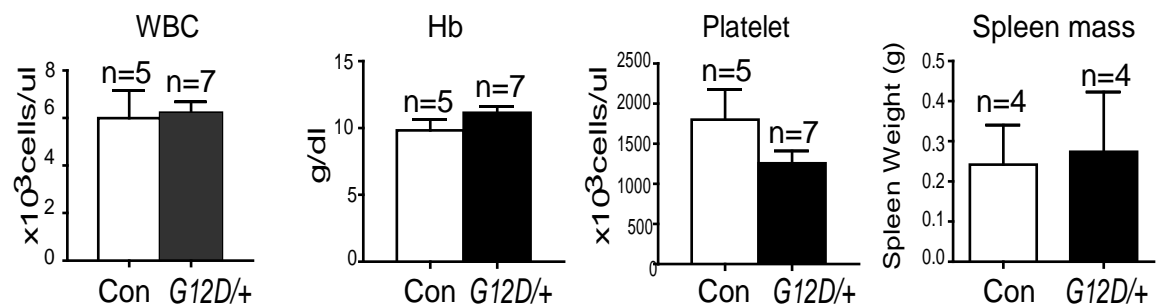
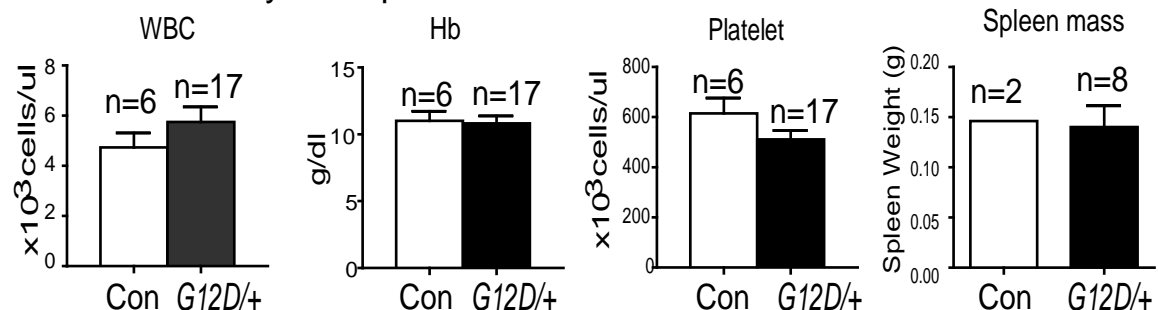
Extended Data Figure 3 | HSCs from *Mx1-cre; Nras^{G12D/+}* mice were not immortalized. A fifth round of serial transplantation of 3×10^6 bone marrow cells from the quaternary recipients of *Nras^{G12D/+}* (*G12D/+*) bone marrow cells shown in Fig. 2c showed that the *Nras^{G12D/+}* HSCs eventually exhausted all of their HSCs and MPPs and were able to only give low levels of lymphoid

reconstitution. Four donor mice from Fig. 2c were transplanted 20 weeks after the fourth round of transplantation into 4 recipients per quaternary donor. The data represent mean \pm s.d. for donor blood cells in the myeloid (*Gr-1⁺* or *Mac-1⁺* cells), B (*B220⁺*), and T (*CD3⁺*) cell lineages.



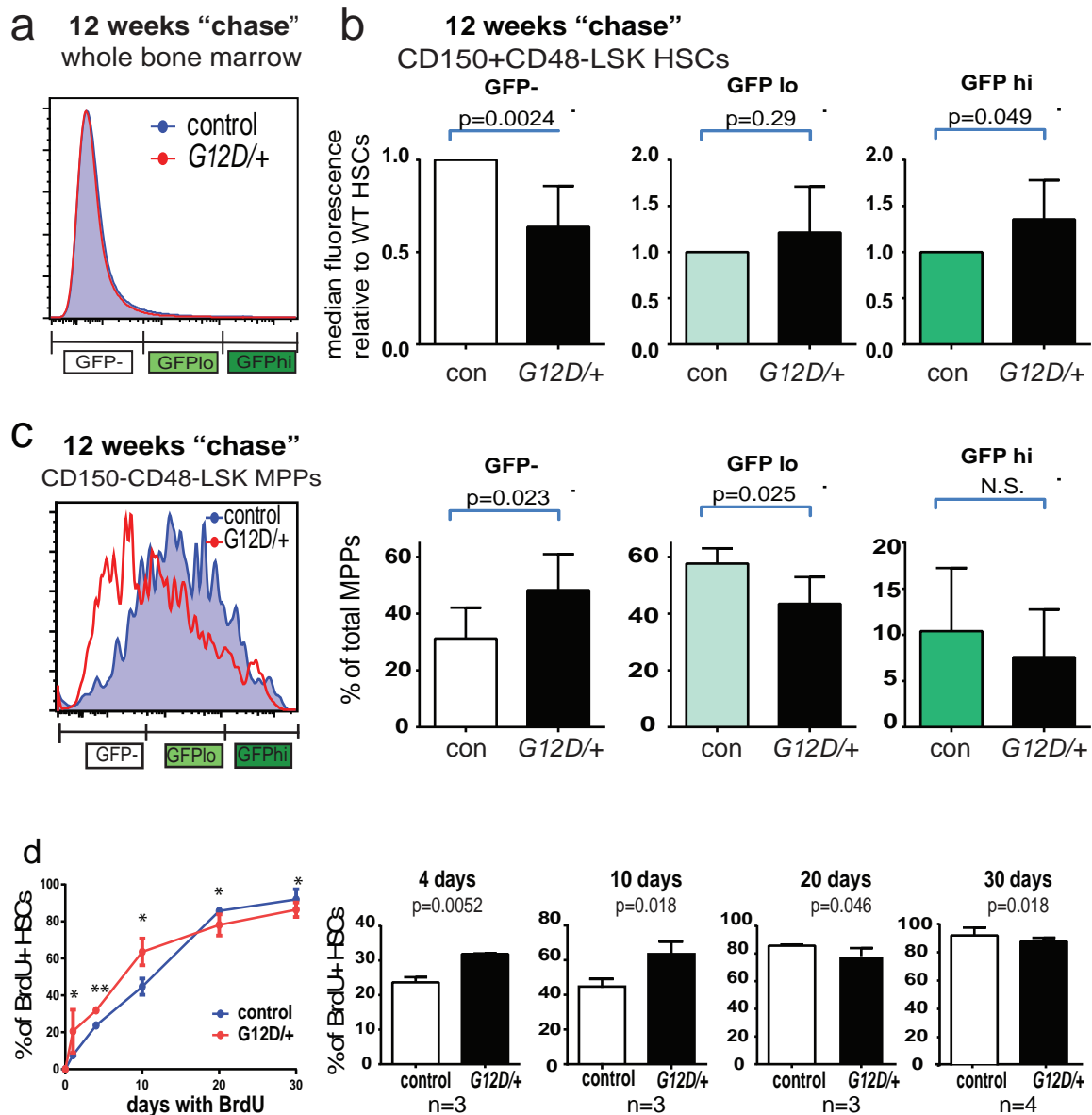
Extended Data Figure 4 | *Nras*^{G12D} (*G12D/+*) expression increased the reconstituting potential of CD150⁺CD48⁺LSK MPPs but did not affect the reconstituting potential of CD150⁺CD48⁺LSK, or CD150⁺CD48⁺LSK progenitors in irradiated mice. a–c, Ten donor MPPs (a), 25 CD150⁺CD48⁺LSK progenitors (b), or 100 CD150⁺CD48⁺LSK progenitors (c) from *Mx1-cre; Nras*^{G12D/+} (*G12D/+*) or littermate control mice at 2 weeks after pIpC treatment were transplanted into irradiated recipient mice along

with 3×10^5 recipient bone marrow cells. Data represent mean \pm s.d. for donor blood cells in the myeloid (Gr-1⁺ or Mac-1⁺ cells), B (B220⁺) and T (CD3⁺) cell lineages. Two-tailed Student's *t*-tests were used to assess statistical significance. None of the time points were significantly different between treatments. The data represent two independent experiments with 4 recipient mice per donor.

a Primary Transplant**b Secondary Transplant****c Tertiary Transplant****d Quaternary Transplant**

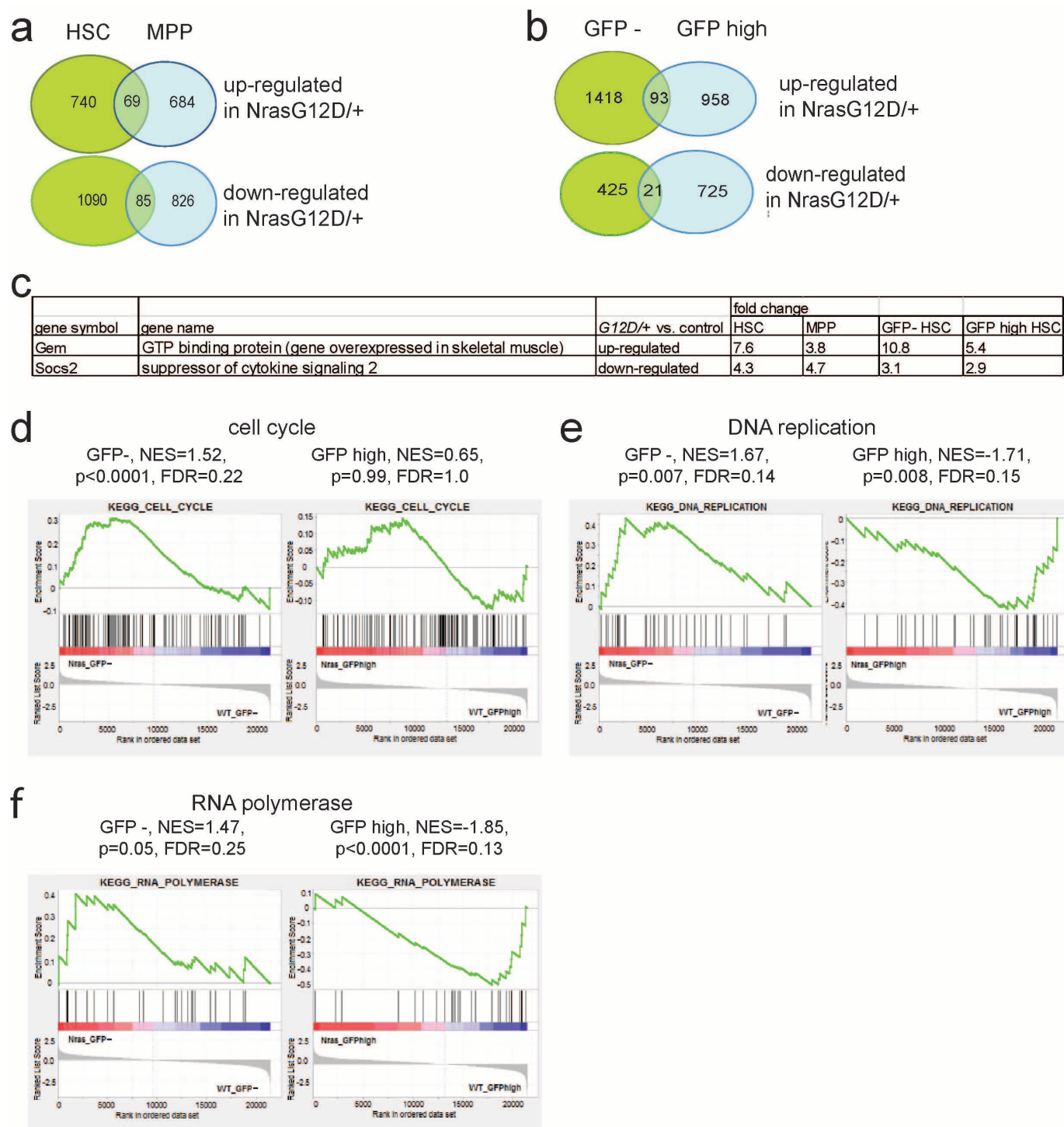
Extended Data Figure 5 | *Nras*^{G12D}-induced changes in HSC function were not associated with the development of leukaemia. a–d, White blood counts (WBC), hemoglobin (Hb) levels, platelet counts and spleen masses for recipient mice from primary transplants (a, from Fig. 1d), secondary transplants (b, from Fig. 2a), tertiary transplants (c, from Fig. 2b) and quaternary transplants (d, from Fig. 2c). In all cases, these blood cell counts were collected from mice after the analysis of blood cell reconstitution was complete (at least 20 weeks after transplantation). The transplanted mice were

observed for a median time of 260 (162–315) days for primary recipient mice, 194 (122–264) days for secondary recipient mice, 224 (176–336) days for tertiary recipient mice, and 280 (279–280) days for quaternary recipient mice. We never observed evidence of leukaemia or MPN by histology in these mice. Across all of the experiments, only two recipients of *Nras*^{G12D/+} cells and two recipients of control cells died spontaneously. Data represent mean \pm s.d. Two-tailed Student's *t*-tests were used to assess statistical significance and none of the comparisons showed significant difference.



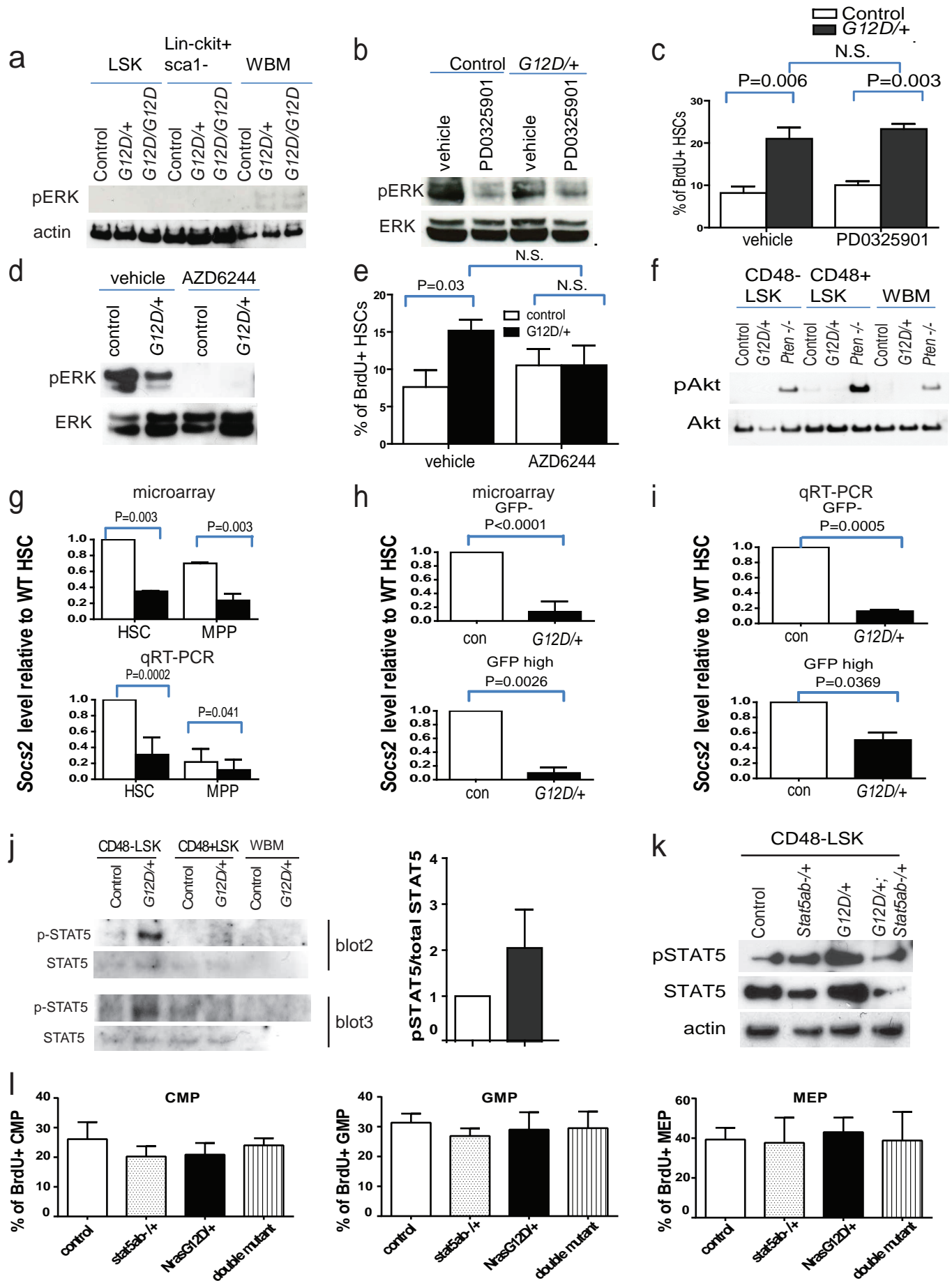
Extended Data Figure 6 | *Nras*^{G12D/+} had a bimodal effect on HSC cycling but increased the rate at which MPPs divide. **a**, Flow cytometric analysis of GFP expression in whole bone marrow cells from *Nras*^{G12D/+} or littermate control mice after 12 weeks of chase without doxycycline. **b**, Median GFP fluorescence intensity of H2B-GFP⁻, H2B-GFP^{lo} and H2B-GFP^{hi} HSCs from wild type and *Nras*^{G12D/+} mice ($n = 8$ mice per genotype). GFP levels in control HSCs were set to one for comparison to relative levels in *Nras*^{G12D/+} HSCs. **c**, *Nras*^{G12D} increased the rate of division by MPPs. Flow cytometric analysis of GFP expression in CD150⁺CD48⁺LSK MPPs from *Mx1-cre*; *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* mice (G12D/+) and littermate controls

(con) after 12 weeks of chase ($n = 8$ mice per genotype). Relative to control MPPs, *Nras*^{G12D/+} MPPs included significantly more H2B-GFP⁻ frequently cycling cells and significantly fewer H2B-GFP^{lo} MPPs ($P < 0.05$ by two-way ANOVA and post hoc pairwise t -tests). **d**, We continuously administered BrdU to *Mx1-cre*; *Nras*^{G12D/+} versus control mice for 1 to 30 days and determined the frequency of BrdU⁺ HSCs (1 day BrdU data are from Fig. 1a). Data represent mean \pm s.d. Two-tailed Student's t -tests were used to assess statistical significance unless stated otherwise. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



Extended Data Figure 7 | Gene expression profiling demonstrates different transcriptional responses to *Nras* activation in quiescent as compared to frequently dividing HSCs. **a**, CD150⁺CD48⁻LSK HSCs and CD150⁺CD48⁻LSK MPPs were isolated from three pairs of *Mx1-cre*; *Nras*^{G12D/+} and littermate controls and gene expression profiling was performed with Affymetrix mouse genome 430 2.0 microarrays. The Venn diagram shows the number of genes that were differentially expressed between *Nras*^{G12D/+} and controls cells within each cell population (fold change ≥ 2).

b, Venn diagram of genes that were differentially expressed between *Nras*^{G12D/+} and control GFP⁻ HSCs and GFP^{hi} HSCs isolated from 3 pairs of *Mx1-cre*; *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* mice and littermate controls (fold change ≥ 2 and P value ≤ 0.05). **c**, Genes that were consistently increased or decreased in expression in response to *Nras* activation in HSCs, MPPs, GFP⁻ HSCs and GFP^{hi} HSCs (fold change ≥ 2 and $P \leq 0.05$ in each cell population). **d-f**, Gene set enrichment analysis (GSEA) of cell cycle genes (**d**), DNA replication genes (**e**) and RNA polymerase genes (**f**).



Extended Data Figure 8 | Nras activation increases STAT5

phosphorylation. **a**, Western blot for phosphorylated ERK (pERK) in LSK stem/progenitor cells, Lin⁻c-kit⁺Sca1⁻ progenitor cells, or whole bone marrow (WBM) cells from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) mice, *Mx1-cre*; *Nras*^{G12D/G12D} (G12D/G12D) mice, or littermate controls 2 weeks after pIpC treatment. **b**, Western blot of pERK and total ERK in 10⁶ uncultured splenocytes from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or control mice after 8 days of treatment with PD0325901 MEK inhibitor or vehicle (blot is representative of four independent experiments). **c**, The frequency of BrdU⁺ CD150⁺CD48⁻ LSK HSCs after a 24-h pulse of BrdU to *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or control mice after 7 days of PD0325901 MEK inhibitor or vehicle (mean ± s.d. from four experiments). **d**, Western blot of pERK and total ERK in 10⁶ uncultured bone marrow cells from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or control mice after 8 days of AZD6244 MEK inhibitor or vehicle (blot is representative of four independent experiments). **e**, The frequency of BrdU⁺ CD150⁺CD48⁻ LSK HSCs after a 24-h pulse of BrdU to *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or control mice after 7 days of AZD6244 MEK inhibitor or vehicle (mean ± s.d. from four experiments). **f**, Western blot for phosphorylated Akt (pAkt) in CD48⁻ LSK HSCs and MPPs, CD48⁺ LSK progenitors, or WBM cells from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) mice, *Mx1-cre*; *Pten*^{fl/fl} (*Pten*^{-/-}) mice, or littermate controls 2 weeks after pIpC treatment. **g**, *Socs2* transcript levels in HSCs and MPPs from *Mx1-cre*; *Nras*^{G12D/+} (G12D/+) or control mice by

microarray analysis (top, *n* = 3) and qRT-PCR (bottom, *n* = 7). **h**, **i**, *Socs2* transcript levels in GFP⁻ and GFP^{hi} HSCs from *Mx1-cre*; *Nras*^{G12D/+}; *Col1A1-H2B-GFP*; *Rosa26-M2-rtTA* mice and littermate controls by microarray (**h**, *n* = 3) and qRT-PCR (**i**, *n* = 3). **j**, Western blotting showed that pSTAT5 levels were significantly increased in CD48⁻ LSK HSCs and MPPs from *Mx1-cre*; *Nras*^{G12D/+} mice as compared to control mice. Left panel shows western blots of pSTAT5 and total STAT5 from two independent experiments. Right panel shows quantification of pSTAT5 levels from western blots from three independent experiments (signals were quantitated using NIH ImageJ software). Blot 1 was shown in Fig. 4e. **k**, Western blot showing that STAT5 levels were reduced in CD48⁻ LSK HSCs/MPPs from *Mx1-cre*; *Stat5ab*^{-/+} or *Mx1-cre*; *Nras*^{G12D/+}; *Stat5a*^{-/+} mice as compared to control and *Mx1-cre*; *Nras*^{G12D/+} mice (blot is representative of four independent experiments). **l**, BrdU incorporation into common myeloid progenitors (CMPs; Lin⁻Sca1⁻c-kit⁺CD34⁺CD16/32⁻), granulocyte macrophage progenitors (GMPs; Lin⁻Sca1⁻c-kit⁺CD34⁺CD16/32⁺), and megakaryocyte erythroid progenitors (MEPs; Lin⁻Sca1⁻c-kit⁺CD34⁻CD16/32⁻) from control, *Mx1-cre*; *Stat5a*^{-/+}, *Mx1-cre*; *Nras*^{G12D/+}, or *Mx1-cre*; *Nras*^{G12D/+}; *Stat5ab*^{-/+} mice after a 2.5-h pulse of BrdU (*n* = 4 mice per treatment). Data represent mean ± s.d. Two-tailed Student's *t*-tests were used to assess statistical significance.

The protein quality control system manages plant defence compound synthesis

Jacob Pollier^{1,2*}, Tessa Moses^{1,2,3,4*}, Miguel González-Guzmán^{1,2†}, Nathan De Geyter^{1,2}, Saskia Lippens^{5,6}, Robin Vanden Bossche^{1,2}, Peter Marhavý^{1,2}, Anna Kremer^{5,6}, Kris Morreel^{1,2}, Christopher J. Guérin^{5,6}, Aldo Tava⁷, Wiesław Oleszek⁸, Johan M. Thevelein^{3,4}, Narciso Campos^{9,10}, Sofie Goormachtig^{1,2} & Alain Goossens^{1,2}

Jasmonates are ubiquitous oxylipin-derived phytohormones that are essential in the regulation of many development, growth and defence processes. Across the plant kingdom, jasmonates act as elicitors of the production of bioactive secondary metabolites that serve in defence against attackers^{1–3}. Knowledge of the conserved jasmonate perception and early signalling machineries is increasing^{3–6}, but the downstream mechanisms that regulate defence metabolism remain largely unknown. Here we show that, in the legume *Medicago truncatula*, jasmonate recruits the endoplasmic-reticulum-associated degradation (ERAD) quality control system to manage the production of triterpene saponins, widespread bioactive compounds that share a biogenic origin with sterols^{7–9}. An ERAD-type RING membrane-anchored E3 ubiquitin ligase is co-expressed with saponin synthesis enzymes to control the activity of 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR), the rate-limiting enzyme in the supply of the ubiquitous terpene precursor isopentenyl diphosphate. Thus, unrestrained bioactive saponin accumulation is prevented and plant development and integrity secured. This control apparatus is equivalent to the ERAD system that regulates sterol synthesis in yeasts and mammals but that uses distinct E3 ubiquitin ligases, of the HMGR degradation 1 (HRD1) type, to direct destruction of HMGR^{10–13}. Hence, the general principles for the management of sterol and triterpene saponin biosynthesis are conserved across eukaryotes but can be controlled by divergent regulatory cues.

To identify regulators of plant triterpene synthesis, we monitored the transcriptome of suspension-cultured *M. truncatula* cells, known to increase accumulation of saponins, one of the main metabolite classes in this species, after elicitation with jasmonates^{14,15}. The expression of 8,462 transcripts was visualized by complementary DNA-amplified fragment length polymorphism (cDNA-AFLP) transcript profiling and 282 methyl jasmonate (MeJA)-responsive tags were identified. The comparable MeJA-induced expression pattern of the genes encoding HMGR, squalene synthase, squalene epoxidase, β -amyrin synthase (BAS) and CYP93E2, enzymes catalysing steps in saponin biosynthesis^{7–9}, indicated co-regulation (Fig. 1a and Extended Data Figs 1 and 2). Several genes corresponding to potential regulatory factors had maximal transcriptional upregulation before or concurrent with that of the saponin genes, including Myc-like basic helix-loop-helix transcription factors and jasmonate ZIM-domain (JAZ) repressor proteins, known elements of the core jasmonate signalling module^{2–6}, as well as a gene (tag MT067 in Fig. 1a) corresponding to a RING membrane-anchored (RMA)-like E3 ubiquitin ligase¹⁰, which we termed makibishi 1 (*MKB1*) (Extended Data Fig. 3). The early MeJA response of *MKB1*

was confirmed in the *M. truncatula* Gene Expression Atlas (MtGEA; <http://bioinfo.noble.org/gene-atlas/>)¹⁶ (Fig. 1b).

To assess *Mkb1* function, we generated transgenic *M. truncatula* hairy roots in which *Mkb1* was overexpressed (*Mkb1*^{OE}) or knocked down (*Mkb1*^{KD}) (Extended Data Fig. 4a–c). *Mkb1*^{KD} roots showed a striking phenotype, in particular when transferred to liquid medium, which caused dissociation of the *Mkb1*^{KD} roots into caltrop-like structures (Fig. 2a), hence the name ‘makibishi’, Japanese for caltrop. No such phenotypes were observed in control or *Mkb1*^{OE} roots (Extended Data Fig. 4a, b). Microscopic analysis revealed severe morphological deficiencies in *Mkb1*^{KD} roots (Fig. 2b and Extended Data Fig. 4b), which resembled the defects of oat (*Avena sativa*) mutants that accumulate incompletely glycosylated forms of the avenacin saponins¹⁷. Serial block-face scanning electron microscopy indicated that cells had an irregular, instead of the normal cylindrical, shape and that intercellular spaces, typical for the root cortex zone, were completely absent in *Mkb1*^{KD} roots (Fig. 2c, d). This might account for the makibishi phenotype as the severe cell enlacing might render the tissue too rigid and lead to ruptures during root development.

To verify a possible correlation with altered metabolism, we performed metabolite profiling by liquid chromatography electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (LC-ESI-FT-ICR-MS)¹⁸. *Mkb1*^{KD} roots were clearly different from control roots, whereas *Mkb1*^{OE} roots showed no significant differences (Fig. 3a and Extended Data Fig. 4d, e). Tens of compounds showed a significantly higher or lower accumulation in *Mkb1*^{KD} roots, including numerous saponins. Most of the upregulated saponins were monoglycosylated compounds, such as 3-O-Glc-medicagenic acid, whereas only higher glycosylated forms, such as soyasaponin I, were represented within the downregulated saponins (Fig. 3b, c, Extended Data Fig. 4f, g and Extended Data Table 1). Analysis of the growth medium of *Mkb1*^{KD} roots revealed the presence of tens of compounds, including the monoglycosylated saponins, whereas in the growth medium of control roots no metabolites were detected (Extended Data Fig. 5a), indicating release of compounds from the *Mkb1*^{KD} roots. Notably, application of aliquots of *Mkb1*^{KD} root culture medium to control roots caused transient tissue loosening (Extended Data Fig. 5b), suggesting that ectopic accumulation of bioactive monoglycosylated saponins^{7–9,19} and/or other metabolites contributes to the makibishi phenotype and is not a mere consequence of the root defects. These findings suggest that silencing of *MKB1* interferes with biosynthesis of saponins, leading to the overaccumulation and release of incompletely glycosylated saponins and correlating with the manifestation of the makibishi phenotype.

¹Department of Plant Systems Biology, VIB, Technologiepark 927, Gent B-9052, Belgium. ²Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, Gent B-9052, Belgium. ³Laboratory of Molecular Cell Biology, Institute of Botany and Microbiology, KU Leuven, Kasteelpark Arenberg 31, Leuven-Heverlee B-3001, Belgium. ⁴Department of Molecular Microbiology, VIB, Kasteelpark Arenberg 31, Leuven-Heverlee B-3001, Belgium. ⁵VIB Bio Imaging Core Gent, VIB, Technologiepark 927, Gent B-9052, Belgium. ⁶Inflammation Research Center, VIB, Technologiepark 927, Gent B-9052, Belgium. ⁷Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Centro di Ricerca per le Produzioni Foraggere e Lattiero Casearie, viale Piacenza 29, Lodi 26900, Italy. ⁸Department of Biochemistry, Institute of Soil Science and Plant Cultivation - State Research Institute, ul. Czarotryskich 8, Pulawy 24-100, Poland. ⁹Departament de Bioquímica i Biologia Molecular, Facultat de Biologia, Universitat de Barcelona, Avda Diagonal 643, Barcelona 08028, Spain. ¹⁰Departament de Genètica Molecular, Centre de Recerca en Agrigenòmica, Consorci CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, Barcelona 08193, Spain. [†]Present address: Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Ingeniero Fausto Elio s/n, Valencia 46022, Spain.

*These authors contributed equally to this work.

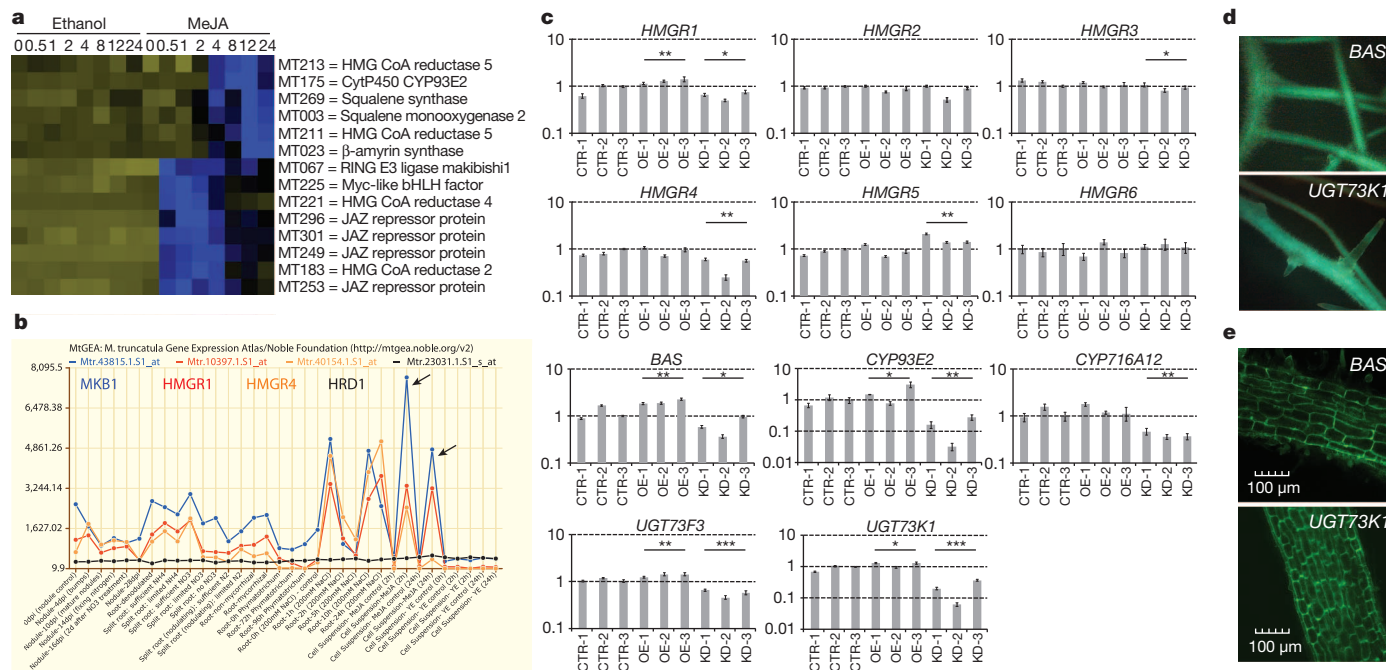


Figure 1 | Expression of *MKB1* and saponin synthesis genes is correlated.

a, Subcluster of the *M. truncatula* transcriptome with genes involved in triterpene biosynthesis or jasmonate signalling. Treatments and time points (in h) are indicated at the top. Blue and yellow boxes reflect transcriptional activation and repression relative to the average expression level, respectively. **b**, Coexpression analysis of *MKB1* (blue), *HMGR1* (red), *HMGR4* (orange) and *HRD1* (black) in *M. truncatula* roots and suspension cells with the MtGEA software¹⁶. Arrows, expression in MeJA-treated suspension cells.

c, Quantitative PCR (qPCR) analysis of saponin genes in control (CTR), *Mkb1*^{OE} (OE) and *Mkb1*^{KD} (KD) roots. *y* axis, expression ratio relative to the normalized transcript levels of CTR line 3 in log scale. Error bars, \pm s.e.m. ($n = 3$). Statistical significance was determined by Student's *t*-test (* $P < 0.1$, ** $P < 0.01$, *** $P < 0.001$). **d**, **e**, Fluorescence (**d**) and confocal (**e**) microscopy analysis of GFP expression driven by the *BAS* and *UGT73K1* promoters in *M. truncatula* hairy roots.

Notably, expression of all hitherto known specific saponin biosynthetic genes (*BAS*, *CYP93E2*, *CYP716A12*, *UGT73F3* and *UGT73K1*) was strongly downregulated in *Mkb1*^{KD} roots (Fig. 1c and Extended Data Fig. 2). Expression of some of the genes corresponding to enzymes catalysing triterpene synthesis up to the oxidosqualene precursor was also reduced, but downregulation was not observed with the putative *M. truncatula* orthologues of genes corresponding to *Arabidopsis thaliana* sterol biosynthesis enzymes (Fig. 1c and Extended Data Fig. 6a–c), indicating that *MKB1* silencing did not affect transcriptional regulation of triterpene biosynthesis in general. Correspondingly, no *Mkb1*^{KD}-specific differences in sterol levels were detected (Extended Data Fig. 6d). Expression of none of the triterpene pathway genes decreased in *Mkb1*^{OE} roots (Fig. 1c and Extended Data Fig. 6b, c), in accordance with the lack of an observable phenotype. These findings point towards the occurrence of a saponin-specific feedback mechanism, probably required to cope with the ectopic accumulation of bioactive monoglycosylated saponins. Verification of the expression of green fluorescent protein (GFP)-reporter constructs, driven by promoters of the *BAS* and *UGT73K1* genes, showed that these genes are ubiquitously expressed in *M. truncatula* hairy roots (Fig. 1d, e), thus precluding that the decrease in their transcript levels is attributable to the developmental defects caused by the makibishi phenotype.

The *Mkb1* protein contains an amino-terminal RING domain and a carboxy-terminal membrane anchor. Accordingly, *Mkb1* possesses self-ubiquitination activity *in vitro* and GFP-tagged *Mkb1* proteins are visible in an endoplasmic-reticulum-reminiscent network pattern in bombarded onion (*Allium cepa*) cells and coincide with a known endoplasmic reticulum protein in yeast (Extended Data Fig. 7a–d). Thus, *Mkb1* corresponds to an active, endoplasmic-reticulum-localized E3 ubiquitin ligase, like its mammalian counterpart RMA1, which is involved in ERAD¹⁰.

Besides protein quality, the ERAD system also controls sterol synthesis in yeasts and mammals through the regulation of HMGR levels.

Yeast does not possess RMA-type proteins and, like mammals, uses HRD1-type ERAD E3 ubiquitin ligases for sterol control^{10–13}. Despite this, the lack of sequence similarity between the different types of E3 ubiquitin ligases and the different membrane topology of the HMGR enzymes from plants, yeasts and mammals (Extended Data Fig. 1b–d), we reasoned that *Mkb1* might survey saponin synthesis in *M. truncatula* by targeting HMGR (Extended Data Fig. 1a). In support of this hypothesis, we observed high expression correlation between several *M. truncatula* *HMGR* genes and *MKB1*, but not the putative *M. truncatula* homologue of yeast *HRD1* (Fig. 1b). Therefore, we checked accumulation of HMGR proteins in *M. truncatula* roots by immunoblot analysis with polyclonal antibodies raised against the conserved catalytic domain of *Arabidopsis* or melon (*Cucumis melo*) HMGR proteins^{20,21}. A small increase in HMGR protein levels was detected in *Mkb1*^{KD} roots as compared to control roots (Extended Data Fig. 8a). Furthermore, we observed that in control roots MeJA application enhanced *HMGR* transcript levels, whereas HMGR protein levels remained stable (Fig. 4a, b). In *Mkb1*^{KD} roots however, HMGR protein levels also increased after MeJA application (Fig. 4b), supporting a role of *Mkb1* in the control of HMGR levels. Unexpectedly, HMGR activity was markedly lower in *Mkb1*^{KD} than in control roots (Extended Data Fig. 8b). Analogous to the effect on the saponin-specific transcripts, we speculate that this is caused by post-translational negative control triggered by the ectopic accumulation of bioactive saponins. A similar inverse correlation between HMGR activity and triterpene levels was observed in transgenic *Taraxacum brevicorniculatum* plants with silenced rubber synthesis, which was postulated to reflect feedback inhibition from oxidosqualene-derived products or precursors²². Multilevel control of HMGR, for example, involving phosphorylation, has been reported in yeast, mammals and plants^{11–13,20,23}.

The relationship between *Mkb1* and HMGR proteins was further examined by three sets of experiments. First, immunoprecipitation assays with tagged *Mkb1* proteins expressed in *M. truncatula* hairy

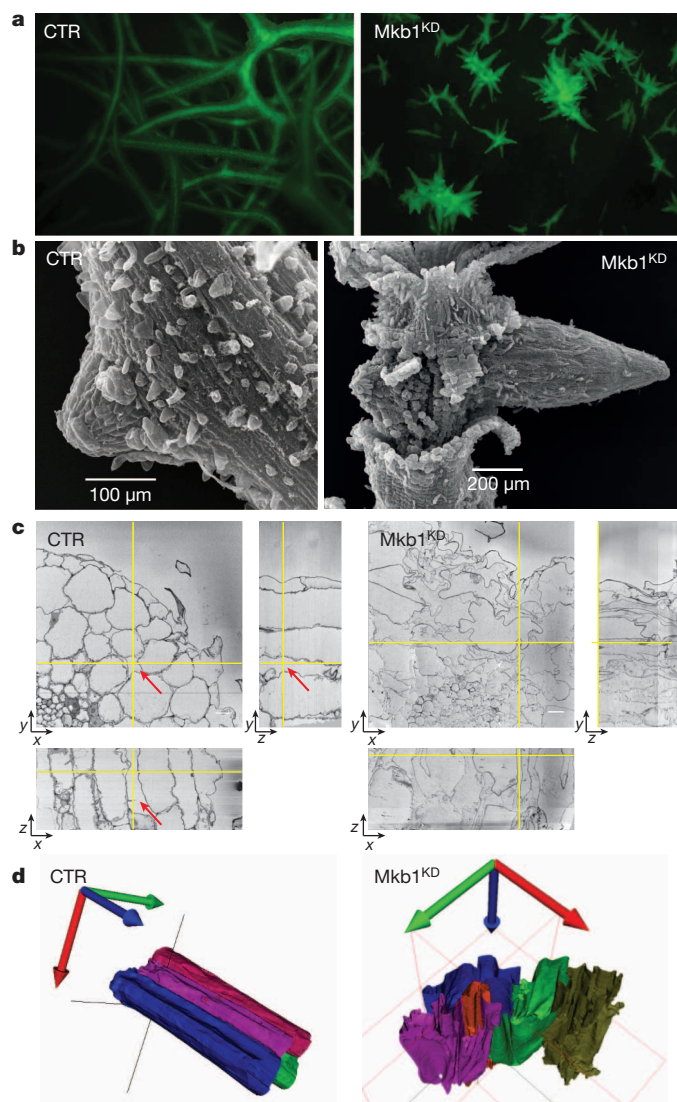


Figure 2 | *MKB1* silencing causes the 'makibishi' phenotype.

a, b, Fluorescence (**a**) and scanning electron (**b**) microscopy analysis of CTR and *Mkb1*^{KD} roots grown in liquid medium. **c, d,** Three-dimensional serial block-face scanning electron microscopy image stacks visualizing the cell structures of CTR and *Mkb1*^{KD} roots grown on solid medium. IMOD, FIJI and Ilastik software were used to generate orthogonal slices (**c**) and three-dimensional reconstructions (**d**). Yellow lines indicate positions of the corresponding orthogonal views. Red arrows indicate the intercellular spaces in CTR roots. Scale bars, 10 μ m.

roots indicated that *Mkb1* and HMGR proteins can physically associate (Extended Data Fig. 7e), probably in an indirect manner as yeast two-hybrid analysis failed to demonstrate direct interaction. This parallels the situation in yeasts and mammals, in which HMGR proteins do not directly interact with HRD-type ERAD ubiquitin ligases but require insulin-induced gene (INSIG)-type proteins as mediators^{11–13}. Second, tagging of firefly luciferase with particular *M. truncatula* HMGR isoforms (*Hmgr1* and *Hmgr3*) converted it into a target of *Mkb1*-mediated protein degradation in transfected tobacco (*Nicotiana tabacum*) protoplasts (Extended Data Fig. 8c). Third, we generated transgenic *M. truncatula* hairy roots overexpressing truncated *Hmgr4* proteins (*tHmgr4*^{OE} roots), devoid of membrane-spanning domains and localizing to the cytosol, by which they are known to be liberated from ERAD control^{11–13}. *tHmgr4*^{OE} roots exhibited a makibishi-like phenotype, accumulated monoglycosylated saponins and showed

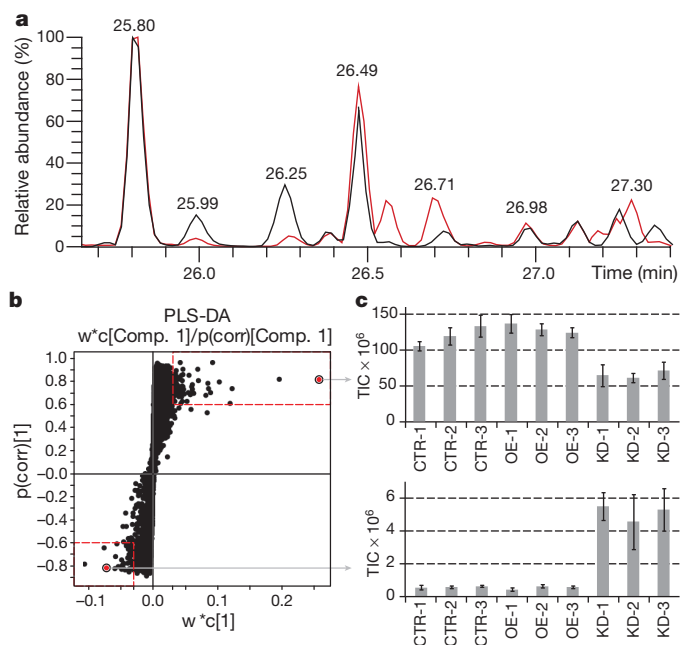


Figure 3 | *MKB1* silencing causes ectopic accumulation of monoglycosylated saponins. **a,** LC-ESI-FT-ICR-MS chromatograms of an extract from CTR (black) and *Mkb1*^{KD} (red) roots. The peak at t_R 26.71 min represents 3-O-Glc-medicagenic acid. **b,** S-plot for correlation ($p(\text{corr})[1]$) and covariance ($w^*c[1]$) derived from partial least squares discriminant analysis (PLS-DA). Metabolites in the bottom left and top right quadrants (marked by dotted red lines) are significantly higher and lower in abundance in the *Mkb1*^{KD} samples, respectively. **c,** Average total ion current (TIC) of the peaks coloured in red in the S-plot and corresponding to soyasaponin I (top) and 3-O-Glc-medicagenic acid (bottom). Error bars, \pm s.e.m. ($n = 4$).

downregulation of saponin gene expression (Fig. 4c, d and Extended Data Figs 9a–d, f–h), demonstrating that loss of *Mkb1* activity and production of 'deregulated' *Hmgr4* proteins cause similar effects. No *tHmgr1*^{OE} lines could be generated; hence we could not determine whether this effect is unique to the *Hmgr4* isoform. Blocking HMGR activity in control roots by lovastatin treatment caused growth inhibition but did not mimic the makibishi phenotype (Extended Data Fig. 9e), confirming that mere loss of HMGR activity cannot account for the *Mkb1*^{KD} effects.

Finally, we demonstrated that *Mkb1* can target yeast HMGR and thereby complement a yeast strain devoid of *Hrd1*, despite the lack of sequence and topology similarity between both the E3 ubiquitin ligases and their native targets (Fig. 4e and Extended Data Figs 1 and 9i). Hence, although *M. truncatula* uses an ERAD system different from those directing sterol-regulated destruction of HMGR enzymes in yeast and mammals, it appears compatible. Mammals and yeasts use sterols or isopentenyl diphosphate-derived non-sterols to regulate HMGR^{24,25}. The divergent sequences of the plant proteins might have allowed the evolution of a plant-specific gateway to the control of HMGR activity, for example, regulated by specific saponin intermediates or involving plant-specific 'mediator' proteins (the INSIG proteins that are conserved between yeast and mammals are not present in plants). Identifying these elements will be key to unravelling the molecular mechanisms that control plant HMGR activity. It is possible that *Mkb1* might manage more endoplasmic-reticulum-localized proteins involved in saponin synthesis, such as cytochrome P450 enzymes, chaperones or regulators. In this way the plant guarantees self-protection from its own weapons and safeguards development and integrity when trying to eliminate attackers.

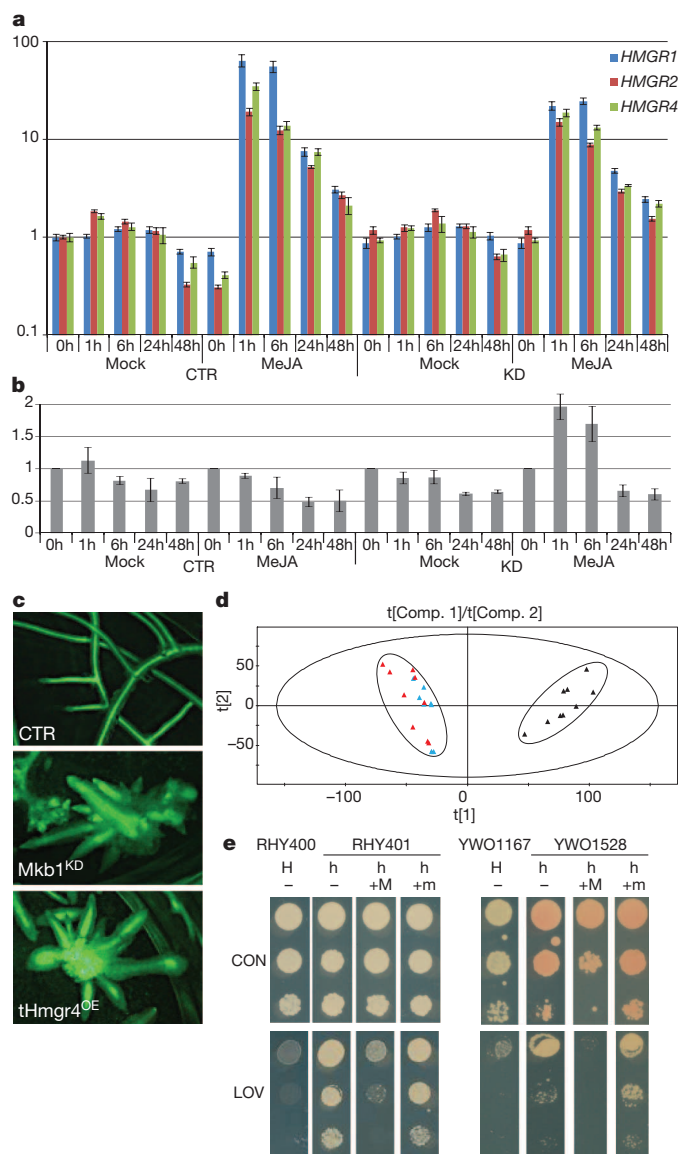


Figure 4 | Mkb1 targets the HMGR enzyme. **a, b**, HMGR expression in mock- or MeJA-treated CTR and Mkb1^{KD} roots. **a**, qRT-PCR analysis of HMGR transcript levels. **b**, Immunoblot analysis of HMGR protein levels. y axis, the ratio relative to the normalized levels of the mock at 0 h. Error bars, \pm s.e.m. ($n = 3$ and 2 , respectively). **c**, Fluorescence microscopy analysis of CTR, Mkb1^{KD} and tHmgr4^{OE} roots grown in liquid medium. **d**, Principal component analysis projecting the first ($t[Comp. 1]$) and second ($t[Comp. 2]$) principal components of the analysis of samples from Mkb1^{KD} (red), tHmgr4^{OE} (blue) and CTR (black) roots. **e**, HRD1 (H) or hrd1 (h) yeasts were transformed with MKB1 (+M) or a ligase-dead version (+m), spotted in a tenfold dilution series on selective synthetic defined medium supplemented (LOV) or not (CON) with lovastatin and grown for 2 days at 30 °C. The empty vector pAG426GPD was used as a control (–). Left and right panels show complementation in the RHY400(H)/RHY401(h) and YWO1167(H)/YWO1528(h) genotypes, respectively.

METHODS SUMMARY

Generation of DNA constructs. Standard molecular biology protocols and Gateway (Invitrogen) technology were followed to obtain expression clones.

Generation and profiling of transgenic *M. truncatula* hairy roots. Transgenic *M. truncatula* (ecotype Jemalong J5) hairy roots were created by *Agrobacterium rhizogenes*-mediated transformation¹⁸. Elicitation, microscopy analysis and metabolite profiling were performed as described^{14,18,26} with modifications.

Mkb1 and HMGR assays. Standard molecular biology protocols were followed to assess localization and activity of Mkb1. HMGR stability and activity were assessed in stably transformed *M. truncatula* hairy roots and in transiently transfected

tobacco protoplasts as described previously^{20,27} with modifications. Immunoprecipitation with tagged *M. truncatula* proteins was performed according to a method developed for *Arabidopsis* proteins²⁸.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 January; accepted 20 September 2013.

Published online 10 November 2013.

- Wasternack, C. Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann. Bot. (Lond.)* **100**, 681–697 (2007).
- Pauwels, L., Inzé, D. & Goossens, A. Jasmonate-inducible gene: what does it mean? *Trends Plant Sci.* **14**, 87–91 (2009).
- De Geyter, N., Gholami, A., Goormachtig, S. & Goossens, A. Transcriptional machineries in jasmonate-elicited plant secondary metabolism. *Trends Plant Sci.* **17**, 349–359 (2012).
- Browse, J. Jasmonate passes muster: a receptor and targets for the defense hormone. *Annu. Rev. Plant Biol.* **60**, 183–205 (2009).
- Memelink, J. Regulation of gene expression by jasmonate hormones. *Phytochemistry* **70**, 1560–1570 (2009).
- Pauwels, L. & Goossens, A. The JAZ proteins: a crucial interface in the jasmonate signaling cascade. *Plant Cell* **23**, 3089–3100 (2011).
- Osborn, A., Goss, R. J. & Field, R. A. The saponins — polar isoprenoids with important and diverse biological activities. *Nat. Prod. Rep.* **28**, 1261–1268 (2011).
- Augustin, J. M., Kuzina, V., Andersen, S. B. & Bak, S. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* **72**, 435–457 (2011).
- Moses, T., Pollier, J., Thevelein, J. M. & Goossens, A. Bioengineering of plant (tri)terpenoids: from metabolic engineering of plants to synthetic biology *in vivo* and *in vitro*. *New Phytol.* **200**, 27–43 (2013).
- Hirsch, C., Gauss, R., Horn, S. C., Neuber, O. & Sommer, T. The ubiquitylation machinery of the endoplasmic reticulum. *Nature* **458**, 453–460 (2009).
- Hampton, R. Y. & Garz, R. M. Protein quality control as a strategy for cellular regulation: lessons from ubiquitin-mediated regulation of the sterol pathway. *Chem. Rev.* **109**, 1561–1574 (2009).
- Jo, Y. & DeBose-Boyd, R. A. Control of cholesterol synthesis through regulated ER-associated degradation of HMG CoA reductase. *Crit. Rev. Biochem. Mol. Biol.* **45**, 185–198 (2010).
- Burg, J. S. & Espenshade, P. J. Regulation of HMG-CoA reductase in mammals and yeast. *Prog. Lipid Res.* **50**, 403–410 (2011).
- Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T. & Dixon, R. A. A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*. *Plant J.* **32**, 1033–1048 (2002).
- Dixon, R. A. & Sumner, L. W. Legume natural products: understanding and manipulating complex pathways for human and animal health. *Plant Physiol.* **131**, 878–885 (2003).
- He, J. *et al.* The *Medicago truncatula* gene expression atlas web server. *BMC Bioinformatics* **10**, 441 (2009).
- Mylona, P. *et al.* Sad3 and sad4 are required for saponin biosynthesis and root development in oat. *Plant Cell* **20**, 201–212 (2008).
- Pollier, J., Morreel, K., Geelen, D. & Goossens, A. Metabolite profiling of triterpene saponins in *Medicago truncatula* hairy roots by liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *J. Nat. Prod.* **74**, 1462–1476 (2011).
- Oleszek, W. Structural specificity of alfalfa (*Medicago sativa*) saponin haemolysis and its impact on two haemolysis-based quantification methods. *J. Sci. Food Agric.* **53**, 477–485 (1990).
- Leivar, P. *et al.* Multilevel control of *Arabidopsis* 3-hydroxy-3-methylglutaryl coenzyme A reductase by protein phosphatase 2A. *Plant Cell* **23**, 1494–1511 (2011).
- Kobayashi, T., Kato-Emori, S., Tomita, K. & Ezura, H. Detection of 3-hydroxy-3-methylglutaryl-coenzyme A reductase protein Cm-HMGR during fruit development in melon (*Cucumis melo* L.). *Theor. Appl. Genet.* **104**, 779–785 (2002).
- Post, J. *et al.* Laticifer-specific cis-prenyltransferase silencing affects the rubber, triterpene, and inulin content of *Taraxacum brevicorniculatum*. *Plant Physiol.* **158**, 1406–1417 (2012).
- Hemmerlin, A. Post-translational events and modifications regulating plant enzymes involved in isoprenoid precursor biosynthesis. *Plant Sci.* **203–204**, 41–54 (2013).
- Sever, N. *et al.* Insig-dependent ubiquitination and degradation of mammalian 3-hydroxy-3-methylglutaryl-CoA reductase stimulated by sterols and geranylgeraniol. *J. Biol. Chem.* **278**, 52479–52490 (2003).
- Garza, R. M., Tran, P. N. & Hampton, R. Y. Geranylgeranyl pyrophosphate is a potent regulator of HRD-dependent 3-hydroxy-3-methylglutaryl-CoA reductase degradation in yeast. *J. Biol. Chem.* **284**, 35368–35380 (2009).
- Wille, S. A. *et al.* Deconstructing complexity: serial block-face electron microscopic analysis of the hippocampal mossy fiber synapse. *J. Neurosci.* **33**, 507–522 (2013).
- De Sutter, V. *et al.* Exploration of jasmonate signalling via automated and standardized transient expression assays in tobacco cells. *Plant J.* **44**, 1065–1076 (2005).

28. Bassard, J. E. *et al.* Protein–protein and protein–membrane associations in the lignin pathway. *Plant Cell* **24**, 4465–4482 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank W. Ardiles-Diaz, S. Carbonelle, R. Dasseville, R. De Rycke and L. Ingelbrecht for technical assistance, and R. Dixon, H. Ezura, R. Hampton, A. Stolz and D. Wolf for providing plant and yeast materials. This research has received funding from the Agency for Innovation by Science and Technology in Flanders ('Strategisch Basisonderzoek' Combiplan project SB0040093), the European Union Seventh Framework Programme FP7/2007–2013 under grant agreement number 222716 – SMARTCELL and the Spanish Ministerio de Economía y Competitividad under grant BFU2011–24208. T.M. and N.D.G. are indebted to the VIB International PhD Fellowship Program and the Agency for Innovation by Science and Technology for predoctoral

fellowships, respectively. J.P. and S.L. are postdoctoral fellows of the Research Foundation Flanders (FWO).

Author Contributions J.P., T.M., M.G.-G., N.D.G., S.L., R.V.B., P.M., A.K., C.J.G., A.T., W.O., N.C. and A.G. performed experiments and analysed the results. J.P., T.M., M.G.-G., N.D.G., S.L., K.M., C.J.G., S.G., N.C. and A.G. designed experiments and analyses. J.P., T.M., J.M.T. and A.G. wrote the manuscript. All authors commented on the results and the manuscript.

Author Information The GenBank EMBL/DDBJ accession number for *MKB1* is JF714982. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.G. (alain.goossens@psb.vib-ugent.be).

METHODS

***M. truncatula* suspension cell culture maintenance and elicitation.** *M. truncatula* cell cultures (provided by R. Dixon) were maintained and elicited as described¹⁴. For elicitation, 7 days after inoculation of 75 ml of a 14-day-old suspension culture into 175 ml fresh medium, cells were treated with 100 μ M MeJA or an equivalent amount of the solvent ethanol as a control. Samples were collected, vacuum filtered and frozen at -80°C .

Transcript profiling. Total RNA from *M. truncatula* cells was prepared with TRIzol (Invitrogen) and reverse transcribed to double-stranded cDNA as described²⁹. After appropriate sample preparation, cDNA-AFLP-based transcript profiling was done with all 128 possible BstYI+1/MseI+2 primer combinations²⁹. Gel image analysis, quantification of band intensities, selection of differentially expressed gene tags, cluster analysis, sequencing and BLAST analysis were carried out as described^{29,30}.

For qRT-PCR, total RNA was extracted with the RNeasy mini kit (Qiagen), and cDNA prepared with SuperScript II Reverse Transcriptase (Invitrogen). Primers were designed with Beacon Designer version 4.0 (Premier Biosoft International). qRT-PCR was carried out with a Lightcycler 480 (Roche) and SYBR Green QPCR Master Mix (Stratagene). For reference genes, 40S ribosomal protein S8 (40S) (TC160725 of the MTGI from TIGR) and translation elongation factor 1 α (ELF1 α) (TC148782 of the MTGI from TIGR) were used. Reactions were done in triplicate and for the relative quantification with multiple reference genes qBase was used³¹.

Generation of DNA constructs. For silencing by means of hairpin RNA interference, the 471-base pair *MKB1* cDNA-AFLP fragment was PCR-amplified and by Gateway recombination cloned into the binary vector pK7GWIGW2D(II)³². The resulting expression clone was transformed into the *Agrobacterium rhizogenes* strain LBA 9402/12 for generation of hairy roots.

To identify the full-length open reading frame (FL-ORF) of *MKB1*, the cDNA-AFLP tag sequence was used for a BLASTN search against the *Medicago truncatula* Gene Index database (<http://compbio.dfci.harvard.edu/tgi/>). The *MKB1* FL-ORF consensus sequence (TC149901; GenBank accession JF714982), the *M. truncatula* *HMGR1*, *HMGR2*, *HMGR3*, *HMGR4* and *HMGR5* sequences (GenBank accessions EU302813, EU302814, EU302815, EU302816 and EU302817, respectively)³³ and the sequences of the *M. truncatula* homologues of *Arabidopsis* *JAZ1* and *CKS1* (GenBank accessions XM_003595306 and XM_003606264, respectively) were PCR-amplified and by Gateway recombination cloned into the entry vector pDONR221. To obtain entry clones with and without stop codon, Gateway primers were designed according to ref. 34. The *MKB1* and *HMGR4* entry vectors were used as a template to amplify truncated versions of the ORFs, as well as to create point mutations with the GeneTailor Site-Directed Mutagenesis system (Invitrogen).

The promoter sequences of *BAS* and *UGT73K1* were retrieved from the *M. truncatula* genome v3.5 (ref. 35) (Medtr4g005190 and Medtr4g031800, respectively). For both promoters, 1,000 bases upstream of the start codon were PCR-amplified and by Gateway recombination cloned into the entry vector pDONR221.

All entry constructs were sequence-verified. For stable overexpression experiments, Gateway recombination was carried out with the pK7WG2D binary vector³², and the resulting clone transformed to *A. rhizogenes*. For transient overexpression in tobacco protoplasts, the ORFs were fused at their C terminus with the firefly luciferase ORF by a fusion PCR and Gateway recombined in the p2GW7 vector³². For localization experiments in onion cells, Gateway recombination was carried out with the pK7WGF2 vector³². For recombinant protein production, the sequences were recombined in the pDEST15 expression vector, and the resulting clones transformed to *E. coli* BL21 (DE3) cells. For the yeast complementation and localization assays, the pAG426GPD vector³⁶ was used as the destination vector. To generate bait proteins for immunoprecipitation, the ORFs were fused either N- or C-terminally to the protein G-Streptavidin (GS) tag by Gateway recombination as described³⁷. For promoter analysis, the *BAS* and *UGT73K1* promoter sequences were put in front of a fusion of the *GFP* and β -glucuronidase (*GUS*) coding sequences in the pKGWFS7 vector³².

Phylogenetic analysis. The protein sequences were aligned with ClustalW and the resulting alignments were manually adjusted. The phylogenetic tree was generated in MEGA 4.0.1 software³⁸, by the neighbour-joining method, and bootstrapping was done with 10,000 replicates. The evolutionary distances were computed with the Poisson correction method, and all positions containing gaps and missing data were eliminated from the data set (complete deletion option).

Generation and phenotypic analysis of transgenic *M. truncatula* hairy roots. *A. rhizogenes*-mediated transformation and cultivation of *M. truncatula* (ecotype Jemalong J5) hairy roots was done according to ref. 18.

Samples for SEM were prepared as described³⁹. In brief, after the first fixation step in 4% paraformaldehyde, 1% glutaraldehyde in 2 mM sodium phosphate buffer, the root samples were fixed in 1% osmium tetroxide solution (Fluka) for 2 h, and subsequently subjected to a dehydration series to 100% ethanol. Next, the root samples were critical-point dried and sputter-coated with gold particles

before they were examined with a JEOL JSM-5600 LV or Zeiss Auriga SEM microscope under an acceleration voltage of 10 kV or 1.5 kV, respectively.

For SBF-SEM²⁶, plant roots were fixed in 0.15 M cacodylate, pH 7.4, 2.5% glutaraldehyde (EMS) and 2% paraformaldehyde (AppliChem) for 2 h. To protect the specimens against mechanical stress, the root tips were dipped in 0.6% (w/v) low melting agarose (Sigma) in PBS. Samples were transferred to fresh fixative and kept overnight at 4°C . The next day, samples were washed five times for 3 min in cold 0.15 M cacodylate buffer. *En bloc* contrast staining was performed by consecutive incubations in heavy-metal-containing solutions. Between these steps samples were washed five times for 3 min in ultra-pure water (UPW). The first staining step was 1-h incubation on ice in 1.5% potassium ferrocyanide and 2% aqueous osmium tetroxide in 0.15 M cacodylate buffer. After washing, the samples were incubated for 20 min in a fresh thiocarbonylhydrazide solution (Sigma) (1% (w/v) in UPW) at room temperature (23°C). The next wash step was followed by incubation in 2% osmium in UPW at room temperature for 30 min and overnight incubation in 2% uranyl acetate (EMS) at 4°C . The following day, Walton's lead aspartate staining was performed for 30 min at 60°C . For this, a 30-mM L-aspartic acid solution was used to freshly dissolve lead nitrate (Sigma) (20 mM, pH 5.5). The solution was filtered after 30-min incubation at 60°C . After final washing steps the samples were dehydrated using ice-cold solutions of 70%, 90% and 100% ethanol (anhydrous), for 10 min each. Resin embedding was done using Durcupan AMC (EMS) by first placing the samples in 50% ethanol/Durcupan overnight, followed by two incubations in 100% Durcupan (8 h and overnight). The next day samples were put in fresh Durcupan solution and placed at 60°C for 48 h. For SBF imaging the resin-embedded root tips were mounted on an aluminium specimen pin (Gatan), using conductive epoxy (Circuit Works) and the root tip facing upward. The specimens were trimmed in a pyramid shape using an ultramicrotome and coated with 5 nm of Pt, in a Quorum sputter coater (Quorum Technologies). The aluminium pins were imaged with a Gatan 3View2 in a Zeiss Merlin SEM, using 1.3 kV and the Gatan Digiscan II ESB detector. For registration of the three-dimensional image stack, IMOD (<http://bio3d.colorado.edu/imod/>) was used. Orthogonal views and linear brightness contrast adjustments were obtained in Fiji (<http://fiji.sc/Fiji>). For segmentation and isosurface rendering Ilastik 0.5 was used (<http://www.ilastik.org>). The data sets were automatically segmented using the seeded watershed algorithm.

Metabolite profiling. *M. truncatula* hairy roots were grown for 21 days in liquid medium. The hairy roots were collected and the medium collected from five biological repeats of three independent transgenic lines per transgene construct. Processing and metabolite extraction from hairy root tissue was performed as described¹⁸.

To remove salts from the samples of the culture medium, 1 ml of medium was brought on a 100-mg Extract-Clean SPE column (Mandel) preconditioned with 1 ml 100% MeOH and 1 ml water acidified with 0.1% (v/v) acetic acid. After washing with 1 ml acidified water, samples were eluted in 1 ml methanol. The methanol eluent was evaporated to dryness under vacuum and the residue dissolved in 200 μ l water for analysis.

LC-ESI-FT-ICR-MS analysis was carried out as described¹⁸. In brief, reversed-phase liquid chromatography was achieved using an Acquity UPLC BEH C18 column (150×2.1 mm, $1.7 \mu\text{m}$; Waters) coupled to a second Acquity UPLC BEH C18 column (100×2.1 mm, $1.7 \mu\text{m}$). The following gradient using water/acetonitrile (99:1, v/v) (solvent A) and acetonitrile/water (99:1, v/v) (solvent B), both acidified with 0.1% (v/v) acetic acid, was run: time 0 min, 5% B; 30 min, 55% B; 35 min, 100% B. The loop size, flow rate and column temperature were 25 μ l, 300 μ l per min, and 80°C , respectively. Negative ionization was obtained using a capillary temperature of 150°C , sheath gas of 25 (arbitrary units), auxiliary gas of 3 (arbitrary units), and a spray voltage of 4.5 kV. Full FT-ICR-MS spectra between m/z 120–1,400 were recorded at a resolution of 100,000. Full FT-MS scans were interchanged with dependent MS² scan events, in which the most abundant ion of the previous FT-MS scan was fragmented, and two dependent MS³ scan events in which the two most abundant daughter ions of the MS² scans were fragmented. The collision energy was set to 35%.

The resulting chromatograms were integrated and aligned with the XCMS package⁴⁰ in R version 2.6.1. with the following parameter values: xcmsSet(fwhm = 6, max = 300, snthresh = 2, mzdiff = 0.1), group(bw = 8, max = 300), retcor(method = loess, family = symmetric). A second grouping was done with the same parameter values. Owing to in-source fragmentation, multiple m/z peaks for each compound were often observed.

The principal component analysis and partial least squares discriminant analysis were performed with the SIMCA-P 11 software package (Umetrics AB) with Pareto-scaled mass spectrometry data. Peaks with an absolute covariance value above 0.03 and an absolute correlation value above 0.6 were considered as significantly different.

For identification of the differential metabolites, MSⁿ spectra were elucidated as described^{18,41,42}. To experimentally validate the annotation of several of the elucidated saponins, representative samples of the medium, Mkb1^{KD} and CTR lines were re-analysed in the presence of standard saponins^{43–47}.

Sterols were extracted from 100 mg of ground roots using methanol, which was dried under vacuum and further extracted with hexane:water (1:1). The hexane phase was dried under vacuum and trimethylsilylated for gas chromatography mass spectrometry (GC-MS) analysis using GC model 6890 and MS model 5973 (Agilent). A 1- μ l aliquot was injected in splitless mode into a VF-5ms capillary column (Varian CP9013, Agilent) and operated at a constant helium flow of 1 ml per min. The injector was set to 280 °C and the oven was programmed at 80 °C for 1 min after injection, ramped to 280 °C at 20 °C per min, held at 280 °C for 30 min, ramped to 320 °C at 20 °C per min, held at 320 °C for 1 min, and finally cooled to 80 °C at 50 °C per min at the end of the run. The MS transfer line was set to 250 °C, the MS ion source to 230 °C, and the quadrupole to 150 °C, throughout. MS spectra were generated by scanning the *m/z* range of 60–800 with a solvent delay of 7.8 min. The areas of the peaks were calculated using the default settings of the AMDIS software (v2.6, NIST).

Ubiquitination assay. Recombinant glutathione S-transferase (GST)–Mkb1 fusion proteins (truncated with or without mutation) were purified according to the manufacturer's instructions with Glutathione Sepharose 4B resin columns (GE Healthcare) from transformed *E. coli* cells, pretreated for 2 h with isopropyl- β -D-1-thiogalactopyranoside (IPTG). A protein refolding step to assure the full ion Zn charge of the GST–Mkb1 fusion proteins was included by incubation with refolding buffer (20 mM HEPES, pH 7.4, 0.02 mM ZnCl₂, 1.5 mM MgCl₂, 150 mM KCl, 0.2 mM EDTA, 20% glycerol, 0.05% Triton X-100) for 1 h at 4 °C.

Ubiquitination reactions were done in a total volume of 30 μ l using 15 μ l of the refolded GST–Mkb1 bound to glutathione resin. The reaction contained 300 ng of GST–Mkb1 fusion protein as E3 ubiquitin ligase, 250 ng of the ubiquitin-activating enzyme (UBE1) from rabbit (BostonBiochem), 400 ng of human recombinant UBCH5A protein (BostonBiochem), and 2 μ g of His₆-Ub from human (BostonBiochem) in ubiquitination buffer (50 mM HEPES, pH 7.4, 2 mM ATP, 5 mM MgCl₂, 2 mM DTT, 0.02 mM ZnCl₂). The ubiquitination reactions were incubated for 1 h at 30 °C and stopped by adding 2 \times Laemmli buffer. Samples were resolved on 8% SDS–PAGE followed by protein immunoblot analysis with RGS/penta/tetra His antibody (Qiagen) and anti-GST (GE Healthcare) antibodies.

Particle bombardment of onion epidermis cells. The constructs for localization were transformed into onion epidermis cells by microparticle bombardment with a PDS-1000/He Biolistic Particle Delivery System (Bio-Rad Laboratories). To this end, 1 mg of 1.6- μ m Gold Microcarriers (Bio-Rad Laboratories) was coated with 5 μ g plasmid DNA according to the manufacturer's instructions. The coated particles were bombarded into onion epidermis slices of approximately 2.5 \times 2.5 cm, placed on solid MS medium (pH 5.8) supplemented with 1% (w/v) sucrose, with 1,100 p.s.i. rupture discs and a vacuum of 0.1 bar. Subsequently, the onion slices were stored in the dark for 24 h at room temperature before analysis by confocal microscopy.

Confocal microscopy. *M. truncatula* hairy roots, bombarded onion slices and transformed BY4742 yeast cells (with an integrated RFP tagged Sec13 protein for ER and Golgi marking)^{48,49} were analysed by confocal microscopy with the FV10 ASW Olympus Confocal with a water immersion 63 \times objective.

Analysis of *M. truncatula* HMGR protein levels and activity. Protein extraction from *M. truncatula* hairy roots was carried out as described²⁰. Determination of HMGR protein levels by immunoblot analysis with polyclonal antibodies raised against the conserved catalytic domain of *Arabidopsis* or melon (*Cucumis melo*) HMGR proteins was performed essentially as described^{20,21}. Determination of HMGR-specific activity was carried out as described²⁰.

Immunoprecipitation assays. Protein extraction from *M. truncatula* hairy roots producing GS-tagged bait proteins was carried out according to a protocol described for *Arabidopsis* cells²⁸. Protein purification and precipitation were performed as described^{28,37} except that precipitation was performed immediately after the elution via the AcTEV digest.

HMGR degradation assays in tobacco protoplasts. Protoplast preparation from tobacco Bright Yellow-2 cells, automated transfection, lysis and firefly luciferase assays were carried out as described²⁷.

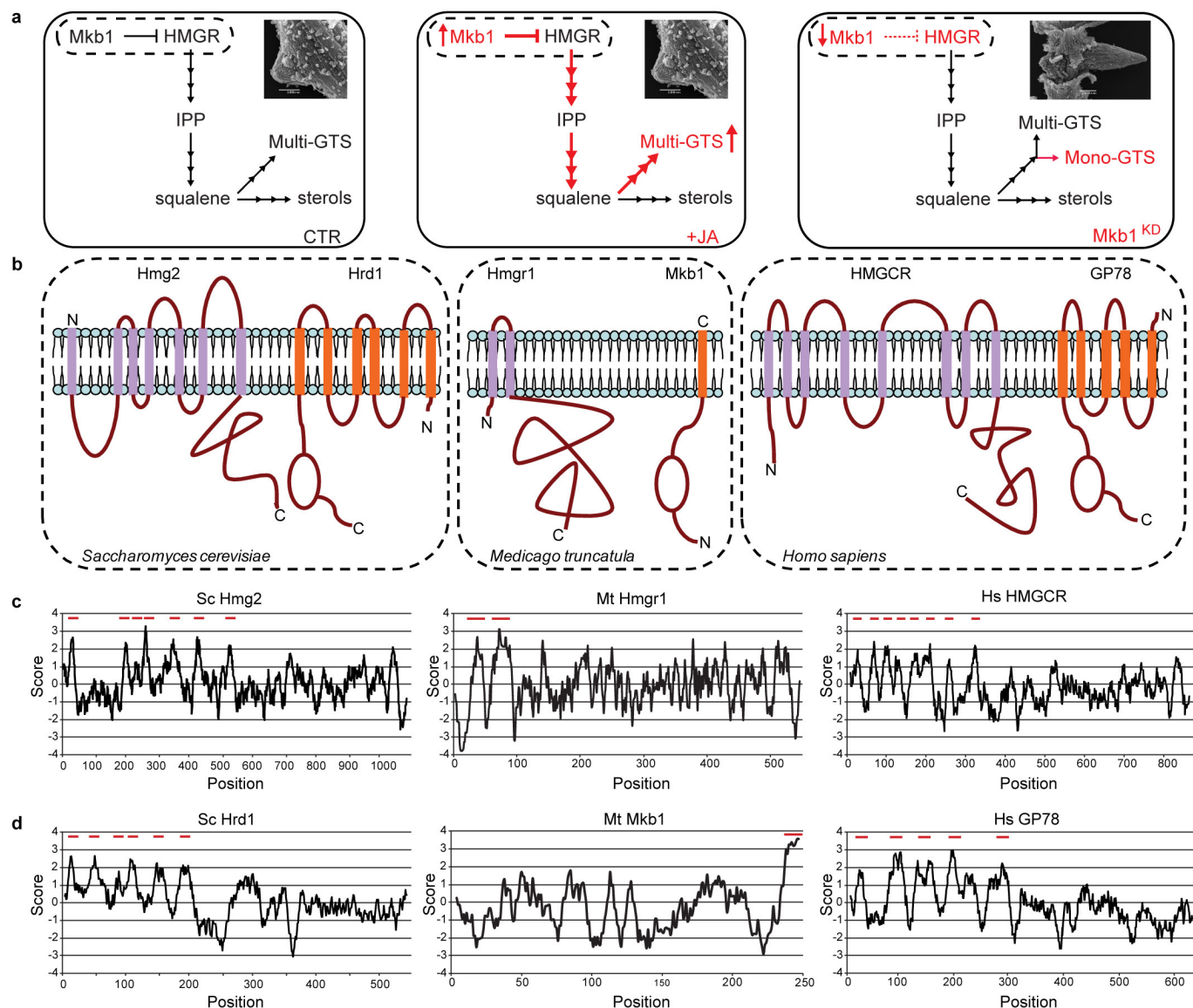
Yeast complementation and protein degradation assays. Two sets of *Saccharomyces cerevisiae* strains were used for the complementation and protein degradation assays, namely strains YWO1167 (*W303 Matx, ura3-1, his3-11,15, leu2-3,112, trp1-1, ade2-1ocre, can1-100, prc1-1, doa10::KanMX*) and its *hrd1* (*Ader3/hrd1::HIS3*) knockout (YWO1528), and RHY400 (*Mata, ade2-101, his3A200, lys2-801, met2, hmg1::LYS2, hmg2::HIS3, ura3-52::6MYC-HMG2*) expressing 6myc–Hmg2

and its *hrd1-1* mutant RHY401 (ref. 50). Transformations were carried out with the high-efficiency lithium acetate/single-stranded carrier DNA/polyethylene glycol method. The transformed yeast strains were selected on minimal medium (2.67% minimal synthetic defined base with 0.077% -Ura dropout supplement; Clontech) supplemented with 30 mg l⁻¹ adenine and methionine.

For the *hrd1* mutant phenotype complementation assays, minimal medium supplemented with 100 μ g ml⁻¹ or 175 μ g ml⁻¹ lovastatin was used for the RHY and YWO strains, respectively. A stock solution of 25 mg ml⁻¹ lovastatin was prepared by the hydrolysis of a 100 mg ml⁻¹ solution in 95% ethanol with 1 N NaOH at 55 °C for 40 min, followed by addition of 1 M Tris-HCl (pH 8.0) and adjustment of pH to 8.0 with 1 N HCl.

The 6myc–Hmg2 level was determined by immunoblotting of whole-cell protein extracts prepared from yeast cells by washing with minimal medium containing 0.1% NaN₃, followed by re-suspension in 100 μ l of SUTE buffer (8 M urea, 1% SDS, 10 mM Tris base, 10 mM EDTA, pH-adjusted to 7.5) containing Complete protease inhibitors (Roche) at pH 6.8. The cells were lysed by vortexing at high speed with acid-washed 0.5 mm glass beads. The lysate was boiled for 10 min at 65 °C after addition of 100 μ l USB buffer (8 M urea, 4% SDS, 0.125 M Tris-HCl, pH 6.8, 10% β -mercaptoethanol, pH-adjusted to 6.8). Ten micrograms of the clear liquid lysate was loaded on SDS–PAGE gels for protein separation, followed by immunoblotting with the 9E10 monoclonal anti-myc antibody.

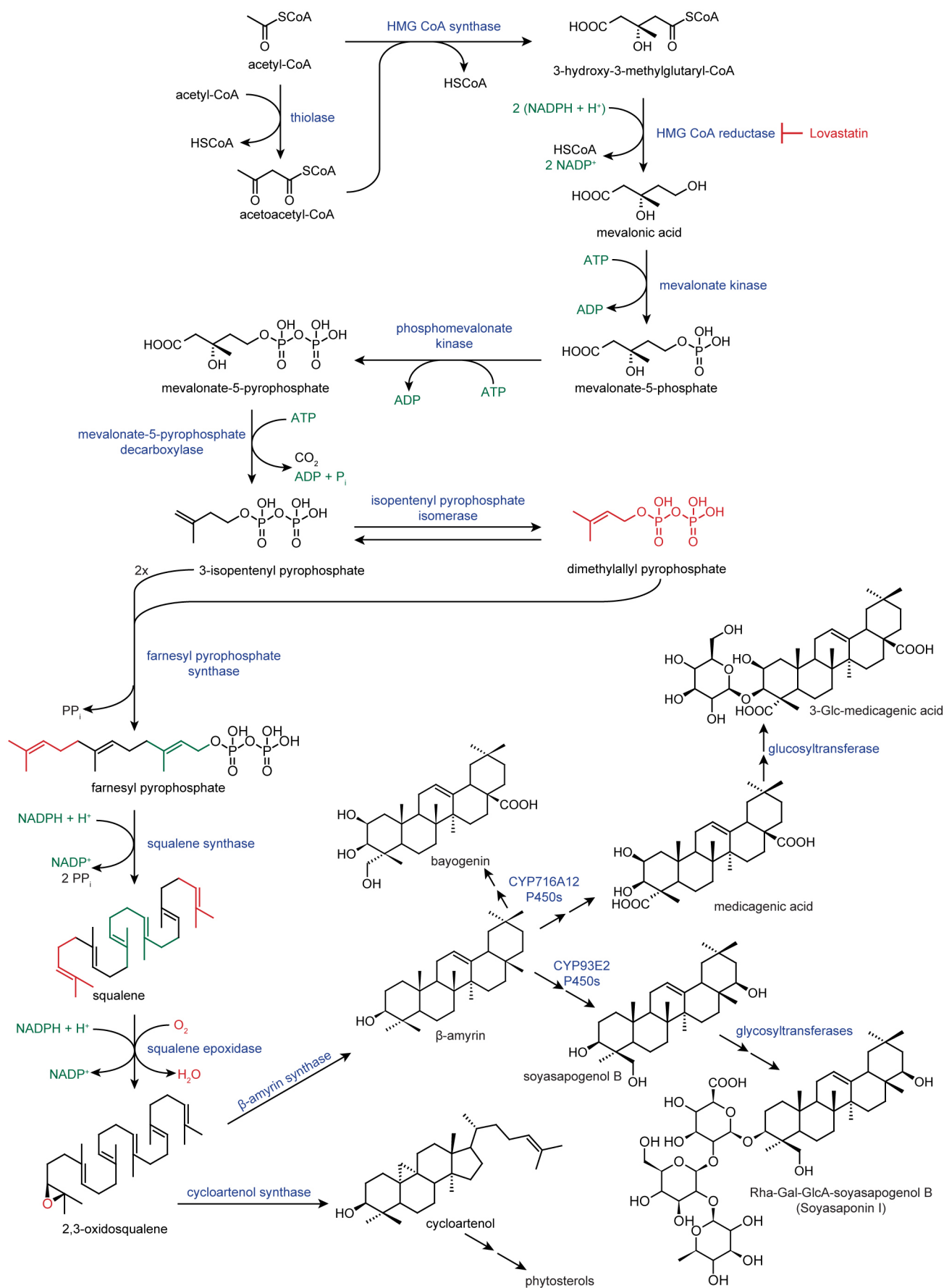
29. Vuylsteke, M., Peleman, J. D. & van Eijk, M. J. T. AFLP-based transcript profiling (cDNA-AFLP) for genome-wide expression analysis. *Nature Protocols* **2**, 1399–1413 (2007).
30. Rischer, H. et al. Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl Acad. Sci. USA* **103**, 5614–5619 (2006).
31. Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. & Vandesompele, J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* **8**, R19 (2007).
32. Karimi, M., Inzé, D. & Depicker, A. GATEWAY™ vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci.* **7**, 193–195 (2002).
33. Kevei, Z. et al. 3-Hydroxy-3-methylglutaryl coenzyme A reductase1 interacts with NORK and is crucial for nodulation in *Medicago truncatula*. *Plant Cell* **19**, 3974–3989 (2007).
34. Underwood, B. A., Vanderhaeghen, R., Whitford, R., Town, C. D. & Hilson, P. Simultaneous high-throughput recombinational cloning of open reading frames in closed and open configurations. *Plant Biotechnol. J.* **4**, 317–324 (2006).
35. Young, N. D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
36. Alberti, S., Gitler, A. D. & Lindquist, S. A suite of Gateway® cloning vectors for high-throughput genetic analysis in *Saccharomyces cerevisiae*. *Yeast* **24**, 913–919 (2007).
37. Van Leene, J. et al. A tandem affinity purification-based technology platform to study the cell cycle interactome in *Arabidopsis thaliana*. *Mol. Cell. Proteomics* **6**, 1226–1238 (2007).
38. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
39. Van Damme, D. et al. Somatic cytokinesis and pollen maturation in *Arabidopsis* depend on TPLATE, which has domains similar to coat proteins. *Plant Cell* **18**, 3502–3518 (2006).
40. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
41. Morreel, K. et al. Genetical metabolomics of flavonoid biosynthesis in *Populus*: a case study. *Plant J.* **47**, 224–237 (2006).
42. Morreel, K. et al. Mass spectrometry-based fragmentation as an identification tool in lignomics. *Anal. Chem.* **82**, 8095–8105 (2010).
43. Oleszek, W. et al. Isolation and identification of alfalfa (*Medicago sativa* L.) root saponins: their activity in relation to a fungal bioassay. *J. Agric. Food Chem.* **38**, 1810–1817 (1990).
44. Tava, A. et al. Triterpenoid glycosides from leaves of *Medicago arborea* L. *J. Agric. Food Chem.* **53**, 9954–9965 (2005).
45. Bialy, Z., Jurzysta, M., Mella, M. & Tava, A. Triterpene saponins from the roots of *Medicago hybrida*. *J. Agric. Food Chem.* **54**, 2520–2526 (2006).
46. Tava, A. et al. New triterpenic saponins from the aerial parts of *Medicago arabica* (L.) Huds. *J. Agric. Food Chem.* **57**, 2826–2835 (2009).
47. Tava, A., Pecetti, L., Romani, M., Mella, M. & Avato, P. Triterpenoid glycosides from the leaves of two cultivars of *Medicago polymorpha* L. *J. Agric. Food Chem.* **59**, 6142–6149 (2011).
48. Knop, M., Finger, A., Braun, T., Hellmuth, K. & Wolf, D. H. Der1, a novel protein specifically required for endoplasmic reticulum degradation in yeast. *EMBO J.* **15**, 753–763 (1996).
49. Huh, W.-K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
50. Hampton, R. Y., Gardner, R. G. & Rine, J. Role of 26S proteasome and HRD genes in the degradation of 3-hydroxy-3-methylglutaryl-CoA reductase, an integral endoplasmic reticulum membrane protein. *Mol. Biol. Cell* **7**, 2029–2044 (1996).



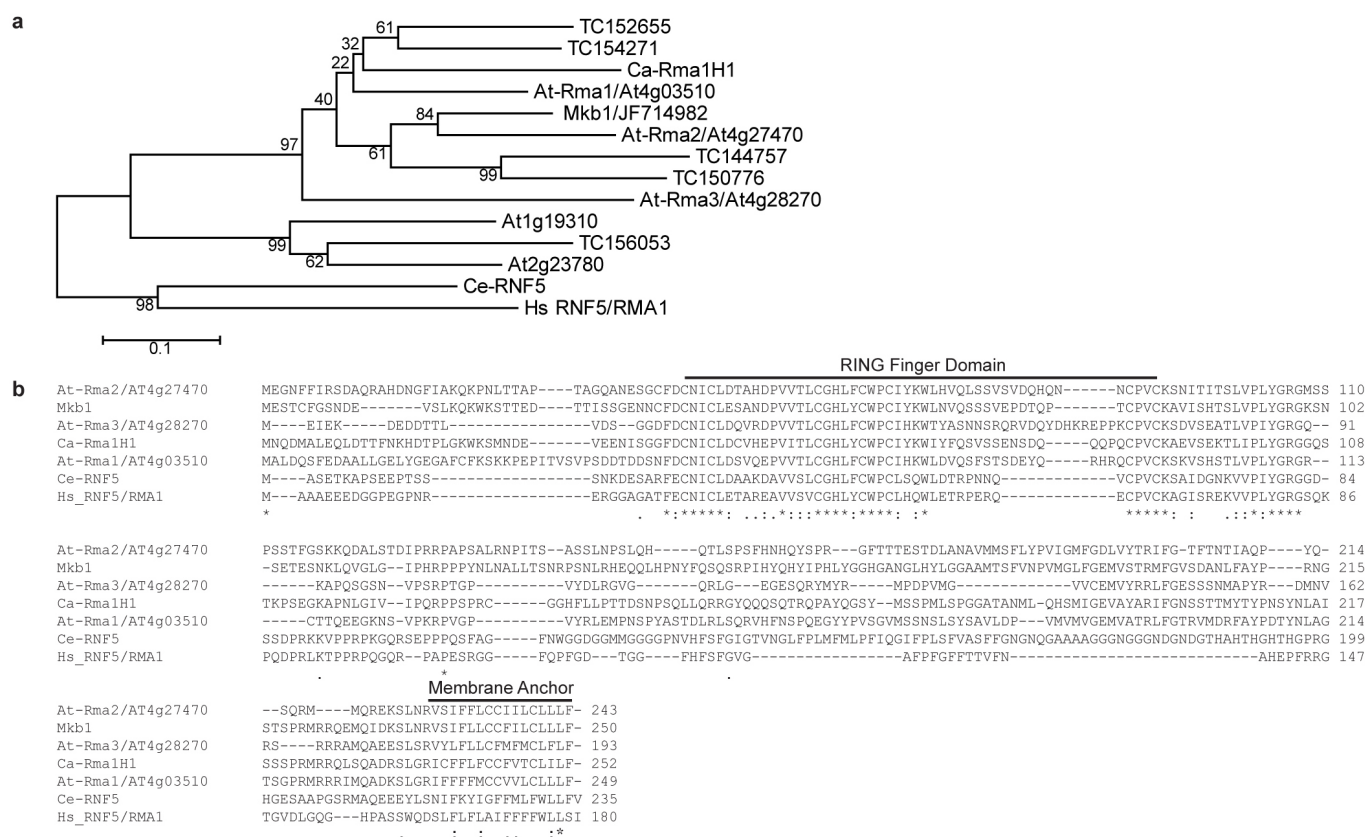
Extended Data Figure 1 | The protein quality control system manages plant defence compound synthesis in the model legume *M. truncatula*.

a, Summarizing schematic. The model depicts three cellular contexts in *M. truncatula* roots in which distinct ERAD-mediated control of HMGR activity occurs and the consequences thereof on root development (inset picture) and triterpene biosynthesis (sterols and glycosylated triterpene saponins (GTS)). The three conditions are (1) control roots cultured in control conditions (CTR; left) with normal ERAD survey of HMGR; (2) control roots cultured in the presence of jasmonate (+JA; middle) with increased triterpene saponin synthesis and increased ERAD activity; and (3) the *Mkb1*^{KD} mutant roots (right) with reduced ERAD control of HMGR activity, leading to accumulation of bioactive monoglycosylated triterpene saponins. Dotted lines

represent the endoplasmic reticulum. Arrows indicate flux through the pathway. Red colours reflect changes in comparison to the CTR condition. IPP, isopentenyl diphosphate. **b**, Schematic overview of the topology of HMGR enzymes and RMA- and HRD-type E3 ubiquitin ligases from yeast, *M. truncatula* and humans. **c**, Kyte–Doolittle hydropathy plot of *S. cerevisiae* Hmg2 (left), *M. truncatula* Hmgr1 (middle) and *H. sapiens* HMGCR (right), with window size 15. **d**, Kyte & Doolittle hydropathy plot of *S. cerevisiae* Hrd1 (left), *M. truncatula* Mkb1 (middle) and *H. sapiens* GP78 (right), with window size 15. Red bars indicate the hydrophobic transmembrane domains. GenBank accession numbers: *H. sapiens*: GP78, Q9UKV5; HMGCR, AAH33692; *M. truncatula*: Mkb1, JF714982; Hmgr1, ABY20972; *S. cerevisiae*: Hmg2, DAA09750; Hrd1, CAA99012.

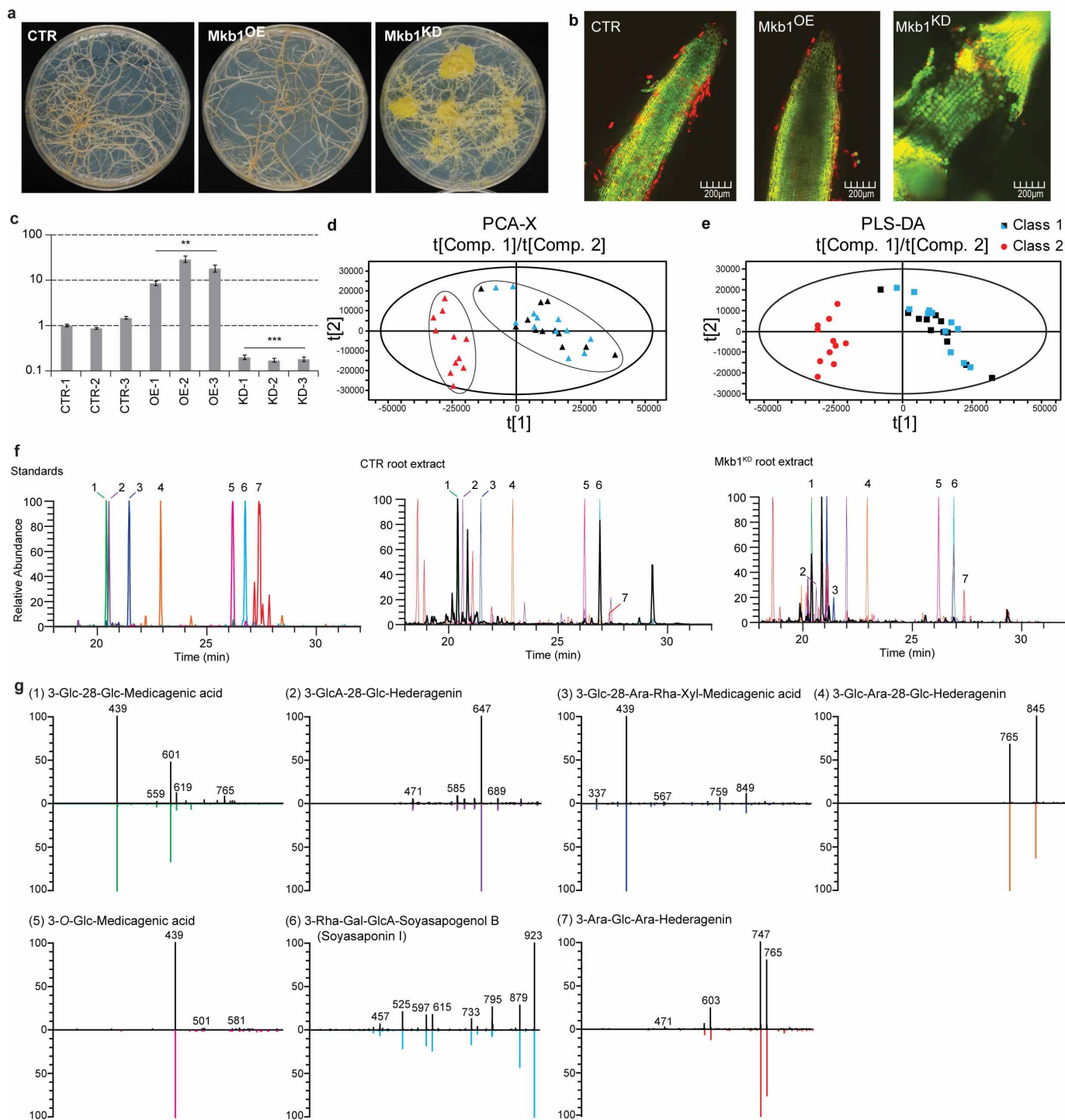


Extended Data Figure 2 | The triterpene saponin biosynthesis pathway in *M. truncatula*. HMG, 3-hydroxy-3-methylglutaryl; P450, cytochrome P450.



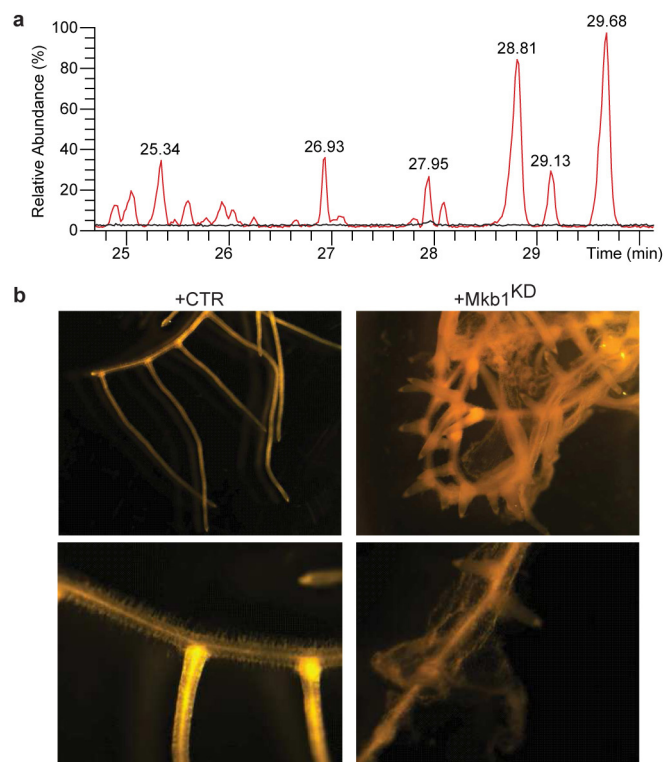
Extended Data Figure 3 | Sequence and structural analysis of eukaryotic RMA proteins. **a**, Phylogenetic analysis of Mkb1 and other RMA-type E3 ubiquitin ligases. The percentage of replicate trees that clustered together in the bootstrap test is shown next to the branches. The scale bar indicates the number of amino acid substitutions per site. *Arabidopsis thaliana* (At), *Capsicum annuum* (Ca), *Caenorhabditis elegans* (Ce) and *Homo sapiens* (Hs) amino acid sequences were retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/>)

genbank/). Amino acid sequences of *M. truncatula* Mkb1 and homologous proteins (prefix TC) were retrieved from the *Medicago truncatula* Gene Index (<http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago>) following BLAST searches. **b**, Comparison of the amino acid sequence of Mkb1 with that of RMA proteins from *A. thaliana*, *C. annuum*, *C. elegans* and *H. sapiens*. Conserved amino acids that are identical in the seven proteins are indicated with an asterisk.

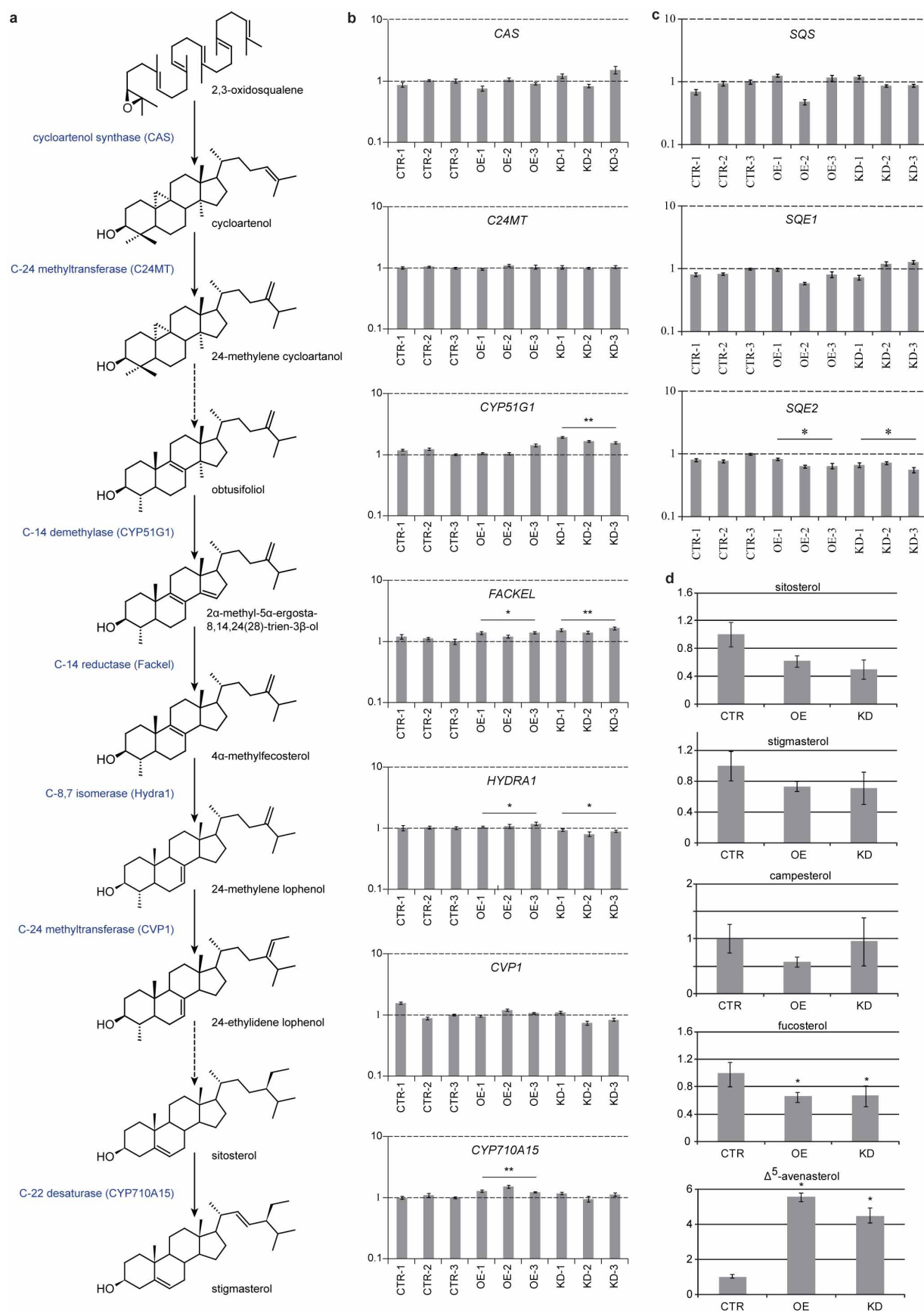


Extended Data Figure 4 | The 'makibishi' phenotype. **a**, CTR, Mkb1^{OE} and Mkb1^{KD} roots grown on solid medium. **b**, Confocal microscopy analysis of CTR, Mkb1^{OE} and Mkb1^{KD} roots grown in liquid medium. **c**, *MKB1* transcript levels in transgenic *M. truncatula* hairy roots. y axis, the expression ratio relative to the normalized transcript levels of CTR line 1 in log scale. Error bars, \pm s.e.m. ($n = 3$). Statistical significance was determined by Student's *t*-test (** $P < 0.01$, *** $P < 0.001$). **d, e**, PCA (**d**) and PLS-DA (**e**) of samples from Mkb1^{KD} (red), Mkb1^{OE} (blue) and CTR (black) roots. **f**, LC-ESI-FT-ICR-MS

chromatograms of seven saponin standards (the identity of which is indicated in **g** and numbered from 1 to 7), an extract of CTR roots, and an extract of Mkb1^{KD} roots (from left to right). The coloured overlay chromatograms depict mass range scans, using a mass window of 0.01 Da, corresponding to the seven standards. **g**, MS² fragmentations of the standards (black, top) compared to the fragmentation of the corresponding peaks in a CTR root extract (coloured, bottom). The numbers correspond to the numbers of the standards depicted in **f**.

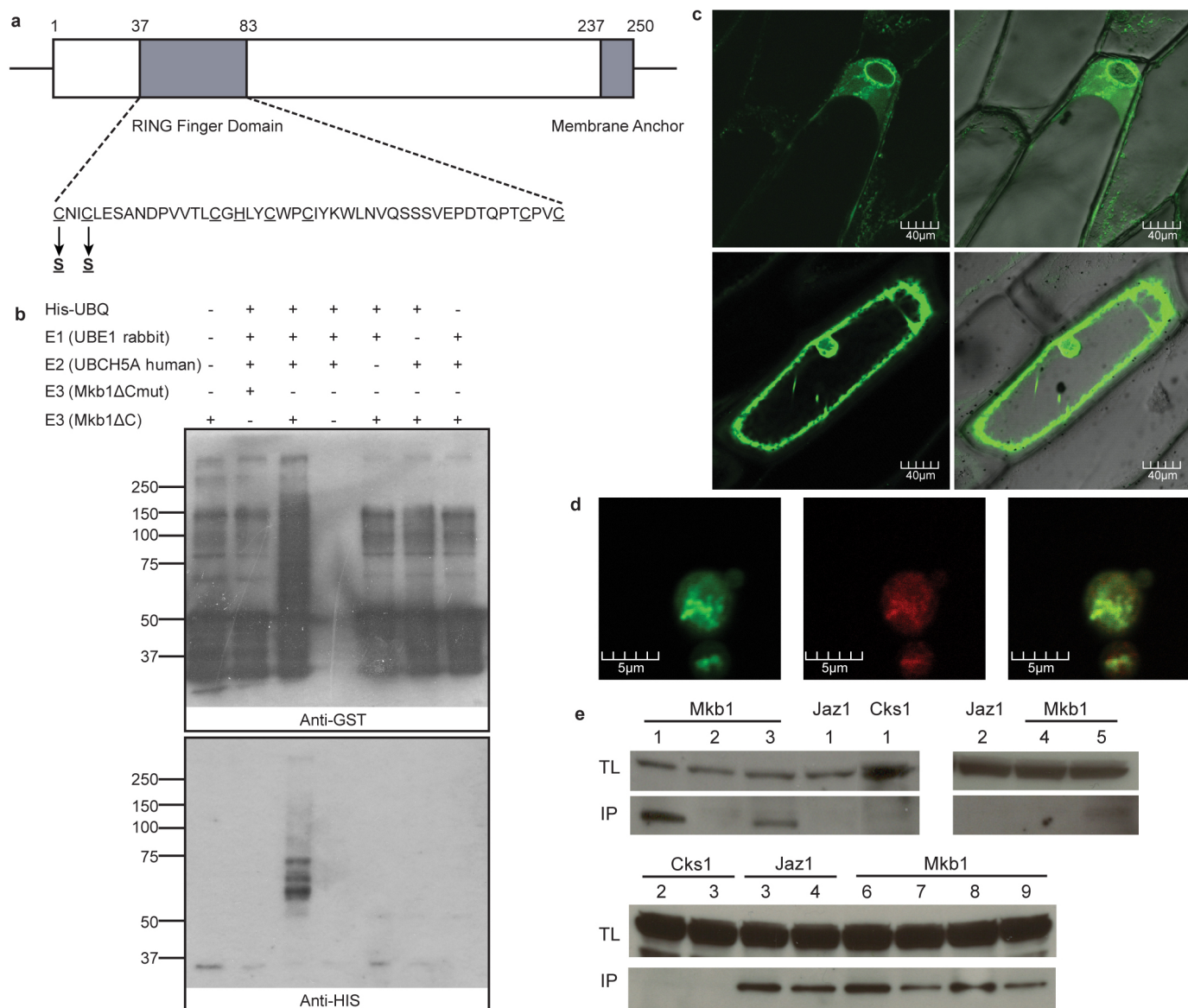


Extended Data Figure 5 | Phenocopy of the Mkb1^{KD} phenotype. **a**, LC-ESI-FT-ICR-MS chromatograms of the medium from CTR (black) and Mkb1^{KD} (red) roots. The peak at t_R 27.95 min represents 3-O-Glc-medicagenic acid. **b**, Light microscopy analysis of CTR hairy roots incubated for 1 week in medium supplemented with medium from CTR (left) or Mkb1^{KD} (right) roots.



Extended Data Figure 6 | Sterol synthesis in transgenic *M. truncatula* hairy roots. **a**, Schematic overview of the sterol biosynthesis pathway. **b**, **c**, qRT-PCR analysis of sterol biosynthetic genes in CTR, Mkb1^{OE} and Mkb1^{KD} roots. **y** axis, the expression ratio relative to the normalized transcript levels of CTR line 3

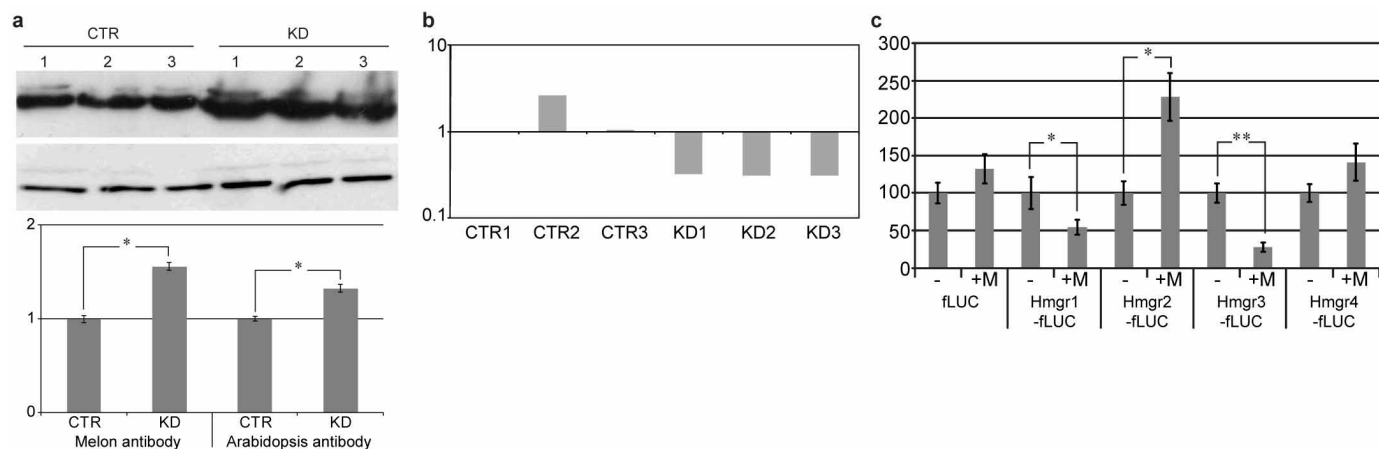
in log scale. SQE, squalene epoxidase; SQS, squalene synthase. **d**, Sterol levels in CTR, Mkb1^{OE} and Mkb1^{KD} roots. **y** axis, sterol accumulation relative to the CTR lines. Error bars, \pm s.e.m. ($n = 3$). Statistical significance was determined by Student's *t*-test (* $P < 0.1$, ** $P < 0.01$).



Extended Data Figure 7 | Mkb1 has auto-ubiquitination activity and is an endoplasmic-reticulum-localized protein that associates with HMGR proteins.

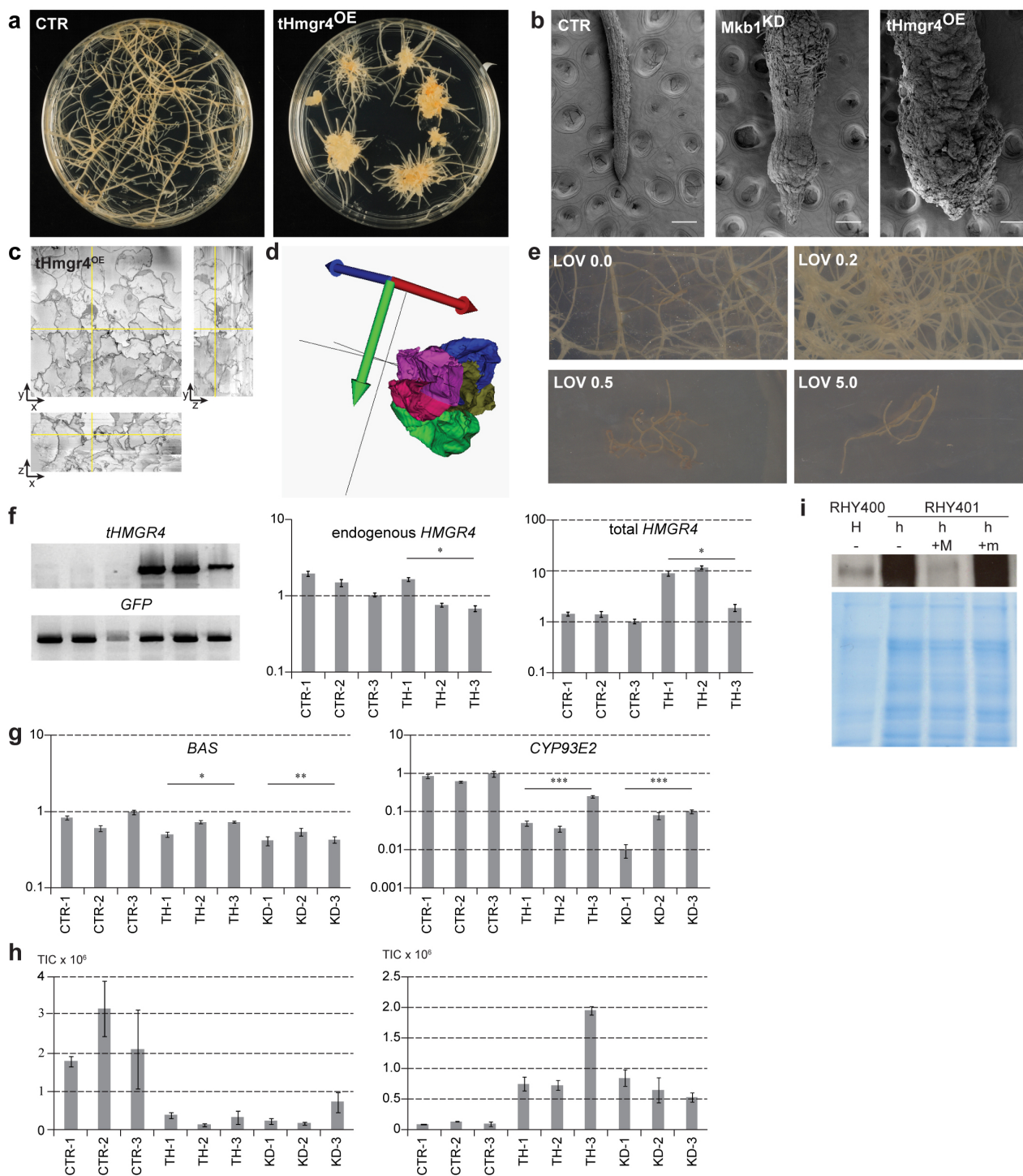
a, Schematic representation of the Mkb1 protein and its domain structure. **b**, *In vitro* auto-ubiquitination assay of Mkb1. The bacterially expressed GST-MKB1 constructs were incubated with ATP in the presence or absence of His-tagged ubiquitin (His-UBQ), E1 (rabbit UBE1) and E2 (human UBCH5A). Samples were resolved by 8% SDS-PAGE, followed by protein immunoblot analysis with anti-GST (top) or anti-His (bottom) antibodies. The recombinant, truncated version of the Mkb1 protein, lacking the membrane anchor domain (Mkb1ΔC), possesses self-ubiquitination activity, whereas a mutated 'ligase-dead' version of the recombinant Mkb1ΔC protein, in which the essential amino acid residues Cys 37 and Cys 40 were substituted by Ser residues, does not. **c**, Subcellular localization of Mkb1 in bombarded onion cells. The pictures show the GFP signal and the GFP-brightfield merged image

(left and right, respectively) of GFP-Mkb1 and GFP-Mkb1ΔC (top and bottom, respectively). The GFP-Mkb1 protein is visible in a network pattern whereas the GFP-Mkb1ΔC protein shows cytosolic localization. **d**, Subcellular localization of Mkb1 in yeast cells. The pictures show the signal of GFP-Mkb1 (left), Sec13-tagged to red fluorescent protein (RFP) (middle), and the merged image (right), respectively. **e**, Total protein lysates (TL, top panels) of *M. truncatula* roots producing GS-tagged versions of Mkb1 or the control proteins Jaz1 (a transcriptional repressor) and Cks1 (a cell cycle control protein) were immunoprecipitated with human IgG Sepharose beads (IP, bottom panels) and subjected to immunoblot analysis with the polyclonal antibodies raised against melon HMGR proteins. In total, association of GS-tagged Mkb1, Jaz1 and Cks1 proteins with HMGR was detected in 7 on 9, 2 on 4, and 0 on 3 independent experiments, respectively.



Extended Data Figure 8 | HMGR levels and activity are altered in *Mkb1*^{KD} lines. **a**, Top, immunoblot analysis with polyclonal antibodies raised against melon (top) and *Arabidopsis* (bottom) HMGR proteins. Bottom, the fold induction in *Mkb1*^{KD} lines relative to the control lines. Error bars, \pm s.e.m. ($n = 3$). Statistical significance was determined by Student's *t*-test (* $P < 0.1$). **b**, Specific HMGR activity in

M. truncatula roots relative to the activity in CTR line 1 in log scale. **c**, The stability of HMGR–firefly luciferase (fLUC) fusion proteins in *MKB1* (+M)-transfected tobacco protoplasts relative to the fLUC value measured in the absence of *MKB1* (–, set at 100%). Error bars, \pm s.e.m. ($n = 24$). Statistical significance was determined by Student's *t*-test (** $P < 0.01$).



Extended Data Figure 9 | Deregulated HMGR activity causes the 'makibishi' phenotype. **a**, CTR and tHmgr4^{OE} hairy roots grown on solid medium. **b**, Scanning electron microscopy analysis of CTR, Mkb1^{KD} and tHmgr4^{OE} roots grown on solid medium. Scale bar, 250 μ m. **c, d**, Three-dimensional serial block-face-scanning electron microscopy image stacks visualizing the cell structures of tHmgr4^{OE} roots grown on solid medium. IMOD, FIJI and Ilastik software were used to generate orthogonal slices (**c**) and three-dimensional reconstructions (**d**). Yellow lines indicate positions of the corresponding orthogonal views. Scale bar, 10 μ m. **e**, Light microscopy analysis of CTR hairy roots maintained for 4 weeks on medium supplemented with increasing amounts of lovastatin (in μ M). **f, g**, Expression analysis of tHmgr4^{OE} lines. **f**, (t)HMGR4 transcript levels in tHmgr4^{OE} roots. The different panels respectively show PCR with reverse transcription (RT-PCR) analysis of the GFP (control) and tHMGR4 transgene transcript levels only (left),

qRT-PCR analysis of the endogenous HMGR4 transcript levels only (middle) and qRT-PCR analysis of total HMGR4 transcript levels (transgene and endogene; right). **g**, qRT-PCR analysis of saponin biosynthetic genes in tHmgr4^{OE} and Mkb1^{KD} roots. y axis, the expression ratio relative to the normalized transcript levels of CTR line 3 in log scale. Error bars, \pm s.e.m. ($n = 3$). Statistical significance was determined by Student's *t*-test (* $P < 0.1$, ** $P < 0.01$, *** $P < 0.001$). **h**, Accumulation of monoglycosylated saponins in tHmgr4^{OE} and Mkb1^{KD} roots. Average total ion current of the peaks corresponding to soyasaponin I (left) and 3-O-Glc-medicagenic acid (right). TH, tHmgr4^{OE} roots. Error bars, \pm s.e.m. ($n = 3$). **i**, Immunoblot analysis for 6myc-tagged Hmg2 and Coomassie blue staining (top and bottom, respectively) of protein extracts from HRD1 (H) or hrd1 (h) yeast cells transformed with MKB1 (+M) or a ligase-dead version (+m). The destination vector pAG426GPD was used as a control (-).

Extended Data Table 1 | LC-ESI-FT-ICR-MS analysis of *M. truncatula* hairy roots

Saponins with identification based on co-analysis relative to an authentic standard showing identical retention and mass data [†]					
Compound Number	S-plot [§]		Compound ID	<i>t_R</i> (min)	[M-H] ⁻
	<i>w</i> [*] <i>c</i> [1]	<i>p</i> (corr)[1]			
Down_01	0,25658	0,81866	3-Rha-Gal-GlcA-Soyasapogenol B (Soyasaponin I)	33,61	941,51560
Down_02	0,12195	0,74033	3-Glc-28-Glc-Medicagenic acid	25,80	825,42776
Down_25	0,03747	0,70874	3-GlcA-28-Glc-Hederagenin	26,49	809,43588
Up_02	-0,07333	-0,81508	3-O-Glc-Medicagenic acid	26,68	663,37701
Down_NS_01 [‡]	0,01221	0,61126	3-Ara-Glc-Ara-Hederagenin	26,27	897,48432
Down_NS_02 [‡]	0,01902	0,84577	3-Glc-28-Ara-Rha-Xyl-Medicagenic acid	26,88	1073,51703
Down_NS_03 [‡]	0,01384	0,67132	3-Glc-Ara-28-Glc-Hederagenin	28,79	927,49638
Saponins with tentative identification based on MS ⁿ data and literature MS data					
Compound Number	S-plot [§]		Compound ID	<i>t_R</i> (min)	[M-H] ⁻
	<i>w</i> [*] <i>c</i> [1]	<i>p</i> (corr)[1]			
Down_05	0,08890	0,70609	Hex-Hex-Hex-Medicagenic acid	25,06	987,48516
Down_08	0,05976	0,75375	Malonyl-Hex-HexA-Bayogenin	25,96	911,43371
Down_09	0,05878	0,69526	dHex-Hex-HexA-Soyasapogenol E	33,00	939,49979
Down_10	0,05766	0,68583	Malonyl-dHex-Hex-HexA-Soyasapogenol B	33,98	1027,51594
Down_12	0,05440	0,70146	Hex-Hex-HexA-Hederagenin	27,10	971,48921
Down_15	0,04526	0,80824	Hex-HexA-Bayogenin	24,89	825,43079
Down_21	0,04360	0,67773	Hex-Hex-HexA-Soyasapogenol E	28,25	955,49360
Down_27	0,03539	0,78195	Malonyl-Hex-dHex-Medicagenic acid	31,06	895,43821
Up_03	-0,05902	-0,81674	Hex-Bayogenin	26,53	649,39774
Up_04	-0,05761	-0,71350	dHex-Hex-Hex-Hederagenin	26,09	941,51536
Up_08	-0,04268	-0,82905	Hex-Hederagenin	28,13	633,40242
Up_12	-0,03721	-0,84694	Malonyl-Hex-Bayogenin	27,17	691,40932
Up_17	-0,03347	-0,76200	Malonyl-Hex-malonyl-Hex-Medicagenic acid	27,26	997,43450
Up_21	-0,03007	-0,75648	Malonyl-Hex-malonyl-Hex-Hederagenin	28,64	967,45877

[†] LC-ESI-FT-ICR-MS chromatograms and MS² fragmentations of the authentic standards compared to the corresponding peaks in *M. truncatula* root extracts are given in Extended Data Fig. 4. Saponins corresponding to five other authentic standards—3-GlcA-28-Ara-Rha-Xyl-medicagenic acid, 3-Ara-Hederagenin, 3-Ara-28-Glc-Bayogenin, 3-Ara-Bayogenin and 3-GlcA-28-Ara-Rha-medicagenic acid—were not detected in *M. truncatula* hairy roots.

[‡] These compounds were less abundant in Mkb1^{HD} roots but below the significance threshold.

[§] Refers to the S-plot in Fig. 3.

^{||} The MSⁿ data and ion intensities for each peak in all analysed extracts are given in Supplementary Table 1.

Histone deacetylase 3 coordinates commensal-bacteria-dependent intestinal homeostasis

Theresa Alenghat^{1,2,3}, Lisa C. Osborne^{1,2}, Steven A. Saenz^{1,2}, Dmytro Kobuley^{1,2}, Carly G. K. Ziegler¹, Shannon E. Mullican^{4,5}, Inchan Choi^{5,6}, Stephanie Grunberg¹, Rohini Sinha¹, Meghan Wynosky-Dolfi^{2,3}, Annelise Snyder³, Paul R. Giacomini^{1,2}, Karen L. Joyce^{1,2}, Tram B. Hoang⁷, Meenakshi Bewtra^{7,8}, Igor E. Brodsky^{2,3}, Gregory F. Sonnenberg^{2,7}, Frederic D. Bushman¹, Kyoung-Jae Won^{5,6}, Mitchell A. Lazar^{4,5,6} & David Artis^{1,2,3}

The development and severity of inflammatory bowel diseases and other chronic inflammatory conditions can be influenced by host genetic and environmental factors, including signals derived from commensal bacteria^{1–6}. However, the mechanisms that integrate these diverse cues remain undefined. Here we demonstrate that mice with an intestinal epithelial cell (IEC)-specific deletion of the epigenome-modifying enzyme histone deacetylase 3 (HDAC3^{ΔIEC} mice) exhibited extensive dysregulation of IEC-intrinsic gene expression, including decreased basal expression of genes associated with antimicrobial defence. Critically, conventionally housed HDAC3^{ΔIEC} mice demonstrated loss of Paneth cells, impaired IEC function and alterations in the composition of intestinal commensal bacteria. In addition, HDAC3^{ΔIEC} mice showed significantly increased susceptibility to intestinal damage and inflammation, indicating that epithelial expression of HDAC3 has a central role in maintaining intestinal homeostasis. Re-derivation of HDAC3^{ΔIEC} mice into germ-free conditions revealed that dysregulated IEC gene expression, Paneth cell homeostasis and intestinal barrier function were largely restored in the absence of commensal bacteria. Although the specific mechanisms through which IEC-intrinsic HDAC3 expression regulates these complex phenotypes remain to be determined, these data indicate that HDAC3 is a critical factor that integrates commensal-bacteria-derived signals to calibrate epithelial cell responses required to establish normal host–commensal relationships and maintain intestinal homeostasis.

Chronic inflammatory diseases, including asthma, allergy, diabetes and inflammatory bowel diseases, are multifactorial diseases that develop as a result of complex gene–environment interactions^{1–6}. In the case of inflammatory bowel disease, genome-wide association studies have identified more than 160 genes or loci that are associated with disease susceptibility⁷. In addition, signals derived from intestinal commensal microbial communities are not only required for normal intestinal function, but also act as environmental cues that influence inflammatory bowel diseases in genetically susceptible hosts^{1–3,8,9}. Intestinal epithelial cells (IECs) function as a crucial cell lineage that integrates microbial signals from the intestinal microenvironment to regulate gene expression and intestinal homeostasis^{10,11}; however, the mechanisms that coordinate these processes remain undefined. Histone deacetylases (HDACs) are epigenome-modifying enzymes that alter gene expression and can be regulated by endogenous factors, dietary components, synthetic inhibitors, and bacteria-derived signals^{12–17}. The class I histone deacetylase HDAC3 alters transcription through histone deacetylation, and may also mediate the activity of other HDACs, deacetylate non-histone targets, and possess enzyme-independent effects^{18–21}. Tissue-specific deletion of HDAC3 in murine models has suggested

critical roles for HDAC3 in complex diseases such as diabetes and heart failure^{22,23}; however, the functional roles of HDAC3 in regulating intestinal homeostasis in the context of health and disease are unknown.

To characterize HDAC3 expression in the intestinal epithelium, intestinal samples from healthy humans and mice were evaluated. HDAC3 protein was expressed in IECs from human and mouse small and large intestine, and immunohistochemistry revealed nuclear localization of HDAC3 in healthy human colonic IECs (Fig. 1a–c). IECs were also isolated from inflammatory bowel disease patients with either Crohn's disease, which commonly targets the terminal ileum, or ulcerative colitis, which is restricted to the large intestine. HDAC3 expression was significantly decreased in IECs isolated from the terminal ileum of Crohn's disease patients (Fig. 1d) and the large intestine of ulcerative colitis patients (Fig. 1e) compared to control patients that had no history of intestinal inflammation, indicating that dysregulated expression of HDAC3 in IECs may be associated with regions of active disease in both forms of inflammatory bowel disease.

To investigate the *in vivo* functions of IEC-intrinsic HDAC3 expression, IEC-specific HDAC3-deficient (HDAC3^{ΔIEC}) mice were generated. HDAC3^{ΔIEC} mice were born at normal Mendelian frequencies, and deletion of HDAC3 was confirmed in IECs (Extended Data Fig. 1a, b). Genome-wide transcriptional profiling on sort-purified live, EpCAM⁺ IECs from the large intestine revealed that *in vivo* deletion of HDAC3 resulted in substantial alterations in IEC-intrinsic gene expression (Extended Data Fig. 1c and Fig. 1f). Most genes that exhibited dysregulated expression were upregulated compared to floxed HDAC3 (HDAC3^{FF}) mice, consistent with a role for HDAC3 in transcriptional repression (Fig. 1f). DAVID (Database for Annotation, Visualization and Integrated Discovery) and gene-set enrichment analyses revealed several HDAC3-dependent pathways in IECs, including those involved in glutathione metabolism, mitochondria, lipid biosynthesis, PPAR signalling, antigen processing and defence response (Fig. 1g and Extended Data Fig. 1d). Altered expression of representative genes in these pathways was confirmed by real-time polymerase chain reaction (PCR) analysis (Fig. 1h). Collectively, these genome-wide analyses implicate a central role for HDAC3 in coordinating a network of IEC-intrinsic transcriptional pathways that regulate multiple cellular processes.

Analyses of histone acetylation in primary IECs from HDAC3^{FF} and HDAC3^{ΔIEC} mice were conducted using genome-wide chromatin immunoprecipitation-sequencing (ChIP-seq) for H3K9Ac, a histone mark that can be a target for HDAC3 at repressed target genes²⁴. ChIP-seq analyses revealed that H3K9Ac levels were significantly increased near genes that were upregulated in IECs isolated from HDAC3^{ΔIEC} mice (Fig. 1i and Extended Data Fig. 1e). The distribution of H3K9Ac at two representative genes, *Scd2* and *Gstp1*, demonstrated multiple sites of

¹Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ²Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ³Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁴Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁵The Institute for Diabetes, Obesity, and Metabolism, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁶Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁷Division of Gastroenterology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁸Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

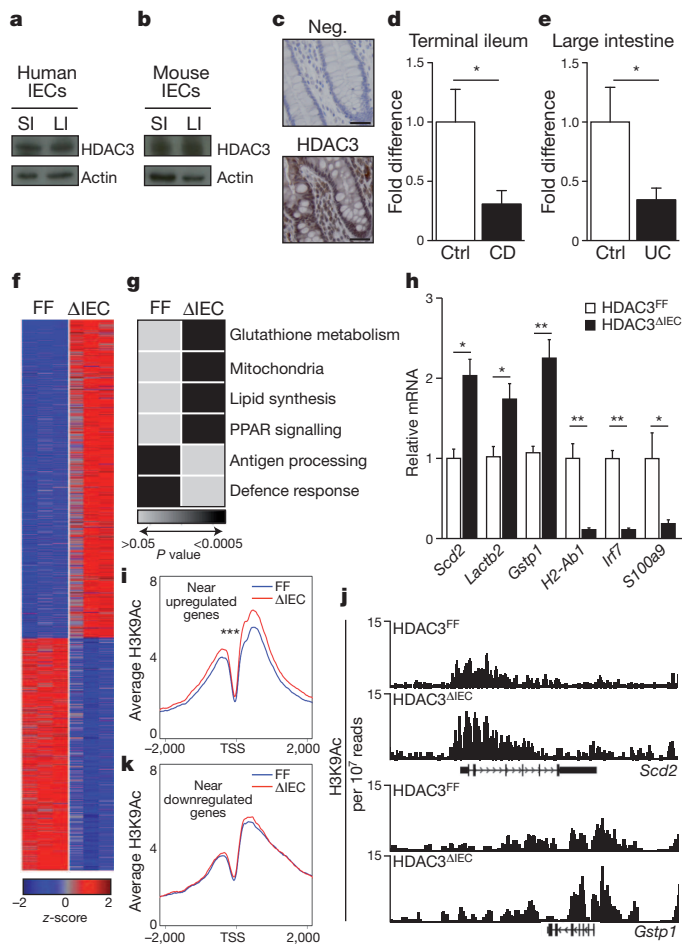


Figure 1 | Decreased expression of HDAC3 in IECs is associated with global alterations in gene expression and histone acetylation. **a, b**, HDAC3 in human (**a**) and mouse (**b**) IECs by western analysis. LI, large intestine; SI, small intestine. **c**, HDAC3 expression in human colon. No primary antibody (Neg.). Scale bars, 25 μ m. **d, e**, HDAC3 mRNA in IECs from terminal ileum (**d**) or large intestine (**e**) of control (Ctrl), Crohn's disease (CD), or ulcerative colitis (UC) patients. * $P < 0.05$, Mann–Whitney U -test (terminal ileum, 8 control, 9 Crohn's disease; large intestine, 10 control, 8 ulcerative colitis). **f**, Expression heat-map in sorted EpCAM⁺ IECs from the large intestine of HDAC3^{FF} (FF) versus HDAC3 ^{Δ IEC} (Δ IEC) mice (fold change > 1.5 , row-normalized Z-score). **g**, Enriched pathways using DAVID. **h**, mRNA in large intestinal IECs. **i**, Average profile of H3K9Ac near upregulated genes in HDAC3 ^{Δ IEC} mice. *** $P = 8.64 \times 10^{-44}$. **j**, Representative distribution of H3K9Ac at select genes from **i, k**. Average profile of H3K9Ac near downregulated genes in HDAC3 ^{Δ IEC} mice. H3K9Ac signals are normalized to reads per 10 million mapped reads. $n = 3$ mice per group. Results are shown as mean \pm s.e.m. ** $P < 0.01$.

acetylation (Fig. 1j), and increased levels of H3K9Ac in multiple genes from HDAC3-deficient IECs were confirmed by ChIP–qPCR (Extended Data Fig. 1f). Collectively, these data indicate that alterations in histone acetylation occur at upregulated genes in IECs deficient in HDAC3. Significant changes in H3K9Ac were not observed near genes for which expression was decreased in IECs from HDAC3 ^{Δ IEC} mice (Fig. 1k), indicating that these genes are unlikely to be direct targets of HDAC3 enzymatic activity. Further analysis will be required to determine whether HDAC3-dependent histone acetylation, deacetylation of non-histone proteins, or enzyme-independent processes are required for specific transcriptional changes observed in IECs isolated from HDAC3 ^{Δ IEC} mice.

Whereas genome-wide analyses demonstrated significant alterations in gene expression in HDAC3 ^{Δ IEC} mice, histological analyses of HDAC3^{FF} versus HDAC3 ^{Δ IEC} mice revealed normal intestinal architecture in HDAC3 ^{Δ IEC} mice (Fig. 2a). However, fundamental alterations

in Paneth cells were observed in HDAC3 ^{Δ IEC} mice, as indicated by significantly decreased numbers of these cells and reduced lysozyme expression (Fig. 2a–c). Paneth cells were observed in postnatal 18-day-old HDAC3 ^{Δ IEC} mice (Extended Data Fig. 2a), indicating that the absence of Paneth cells in adult HDAC3 ^{Δ IEC} mice did not reflect a primary intrinsic developmental defect. Instead, active caspase-3 staining demonstrated elevated cell death in crypts of adult HDAC3 ^{Δ IEC} mice (Extended Data Fig. 2b), and electron microscopy revealed the presence of degenerating organelle membranes and loss of granules in Paneth cells of adult HDAC3 ^{Δ IEC} mice (Extended Data Fig. 2c). Although inhibition of Paneth cell differentiation in adult mice may also occur, these data indicate that impaired Paneth cell survival contributes to altered Paneth cell homeostasis in adult HDAC3 ^{Δ IEC} mice. Furthermore, evaluation of the large intestine revealed that HDAC3 ^{Δ IEC} mice exhibited crypt elongation (Fig. 2d, e) and more extensive Ki67 staining within IECs (Extended Data Fig. 2d), indicating that there is increased IEC proliferation in HDAC3 ^{Δ IEC} mice. Collectively, these findings identify that HDAC3 regulates IEC homeostasis throughout the intestine.

We sought to test whether the alterations in IEC homeostasis observed in HDAC3 ^{Δ IEC} mice were associated with alterations in intestinal barrier function. Naive HDAC3 ^{Δ IEC} mice exhibited increased faecal albumin (Fig. 2f), increased plasma levels of fluorescein isothiocyanate (FITC) after oral administration of FITC-dextran (Extended Data Fig. 3a), and increased lipopolysaccharide (LPS) in mesenteric lymph nodes (Fig. 2g) compared to HDAC3^{FF} mice, indicating that HDAC3 ^{Δ IEC} mice exhibit impaired intestinal barrier function and bacterial translocation. Consistent with the lack of Paneth cells, HDAC3 ^{Δ IEC} mice also exhibited impaired crypt bactericidal activity (Extended Data Fig. 3b) and increased susceptibility to oral *Listeria monocytogenes* infection (Extended Data Fig. 3c, d). Furthermore, as HDAC3 ^{Δ IEC} mice matured, they demonstrated an increased prevalence of rectal prolapse (Fig. 2h and Extended Data Fig. 3e), and colons from these mice exhibited increased inflammation (Extended Data Fig. 3f, g) and elevated disease score (Extended Data Fig. 3h). Collectively, these findings demonstrate that loss of HDAC3 expression in IECs results in impaired barrier function and development of spontaneous intestinal inflammation.

To assess the significance of IEC-intrinsic HDAC3 expression in the context of intestinal damage and inflammation, mice were treated with dextran sodium sulphate (DSS) for 5 days. Whereas HDAC3^{FF} mice were minimally affected, HDAC3 ^{Δ IEC} mice showed profound weight loss (Fig. 3a), increased disease severity (Fig. 3b), colonic shortening (Fig. 3c, d), increased infiltration of neutrophils and macrophages in the intestine (Fig. 3e), and extensive intestinal ulceration, loss of crypt architecture, oedema and inflammation (Fig. 3f), indicating that IEC-intrinsic HDAC3 expression is critical for limiting DSS-induced intestinal damage and inflammation. In addition to IECs, macrophages provide an important link between the microbiota and intestinal homeostasis^{25,26}. Notably, unlike HDAC3 ^{Δ IEC} mice, exposure of HDAC3 ^{Δ LysM} mice²⁴ to DSS did not result in significant weight loss, disease, or intestinal inflammation (Extended Data Fig. 4a–d), indicating that IEC-intrinsic, but not LysM-expressing myeloid-cell-intrinsic, expression of HDAC3 is critical in maintaining intestinal homeostasis and limiting DSS-induced inflammation.

To test how IEC-intrinsic HDAC3 expression functions in adult mice and to eliminate potential developmental effects of constitutive deletion, we generated an inducible tamoxifen-dependent IEC-specific HDAC3 knockout mouse model (HDAC3 ^{Δ IEC-IND}) (Extended Data Fig. 5a) in which depletion of HDAC3 in IECs could be detected after 5 days of tamoxifen treatment (Extended Data Fig. 5b, c). Repeated administration of tamoxifen resulted in decreased Paneth cells and increased caspase-3 staining in the ileal crypts of HDAC3 ^{Δ IEC-IND} mice (Extended Data Fig. 5d), indicating that Paneth cells in adult HDAC3 ^{Δ IEC-IND} mice exhibit altered homeostasis, similar to mice with constitutive deletion of HDAC3 (Extended Data Fig. 2b, c).

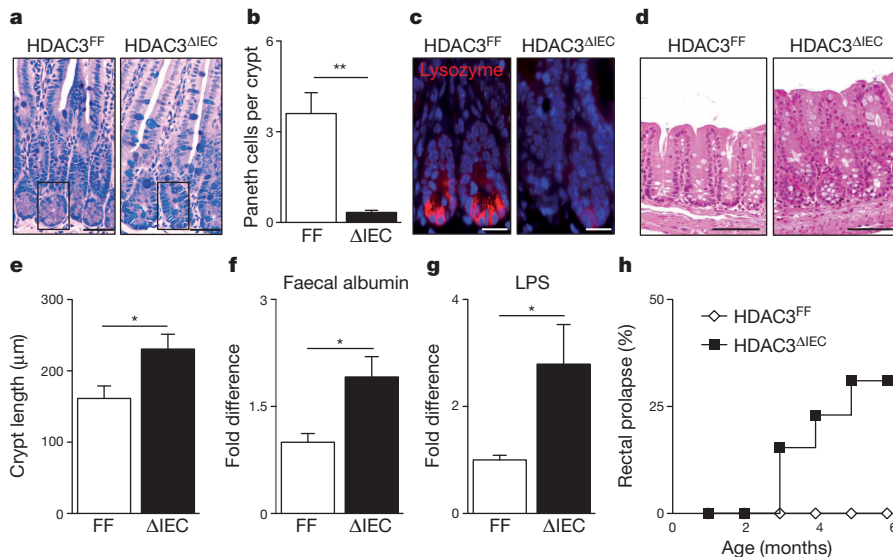


Figure 2 | IEC-intrinsic HDAC3 expression regulates intestinal homeostasis. **a**, Periodic acid-Schiff (PAS)/Alcian-blue-stained ileal sections. The box surrounds Paneth cells. Scale bars, 50 μm. **b**, Paneth cells per crypt. **c**, Immunofluorescent staining of lysozyme (pink) and nuclei (blue). Scale bars, 10 μm. **d**, Colonic sections. Scale bars, 100 μm. **e**, Crypt length in colon. $n = 3$ mice per group. **f**, Albumin from faecal samples of HDAC3^{FF} ($n = 8$) and HDAC3^{ΔIEC} ($n = 6$) mice. **g**, LPS levels in mesenteric lymph node. $n = 8$ mice per group. **h**, Development of rectal prolapse in HDAC3^{ΔIEC} mice ($n = 13$). Data are representative of 2–3 independent experiments. Results are shown as mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$.

Furthermore, increased faecal albumin (Extended Data Fig. 5e) and FITC-dextran permeability (Extended Data Fig. 5f) were observed after deletion of HDAC3, demonstrating that IEC-intrinsic expression of HDAC3 dynamically regulates intestinal barrier function in the adult intestine. Tamoxifen-treated HDAC3^{ΔIEC-IND} mice subjected to DSS exhibited profound weight loss (Fig. 3g), increased disease severity (Fig. 3h), exacerbated colonic shortening (Extended Data

Fig. 6a, b), increased inflammatory cell infiltrates (Extended Data Fig. 6c), and histological lesions in both the large and small intestine (Extended Data Fig. 6d). Collectively, these results identify a critical role for HDAC3 in actively regulating IEC function and tissue homeostasis in adult mice.

Dysregulation of intestinal homeostasis and susceptibility to intestinal inflammation are often associated with alterations in commensal bacterial populations^{27–30}. Therefore, pyrosequencing was used to interrogate temporal and spatial differences in microbial diversity within intestinal bacterial communities. Bacterial communities differed significantly between faecal and intestinal samples of HDAC3^{FF} and HDAC3^{ΔIEC} littermate mice, whereas mice within the same genotype had a more similar bacterial composition (Fig. 3i–l and Extended Data Fig. 7a–c). Most notably, HDAC3^{ΔIEC} mice consistently exhibited increased levels of Proteobacteria (Fig. 3j and Extended Data Fig. 7b, c). Furthermore, analyses of HDAC3^{ΔIEC-IND} mice revealed that alterations in the composition of intestinal bacterial communities occurred after deletion of IEC-intrinsic HDAC3 in adult mice (Extended Data Fig. 7d, e). Recent studies have demonstrated that the colitogenic activity of specific microbial communities can result in exacerbated susceptibility to colitis^{27–29}. However, unlike HDAC3^{ΔIEC} mice, wild-type mice that were cross-fostered (Fig. 3m) or co-housed (Fig. 3n) with HDAC3^{ΔIEC} mice did not demonstrate increased susceptibility to DSS (Fig. 3m, n). Furthermore, germ-free wild-type mice colonized with

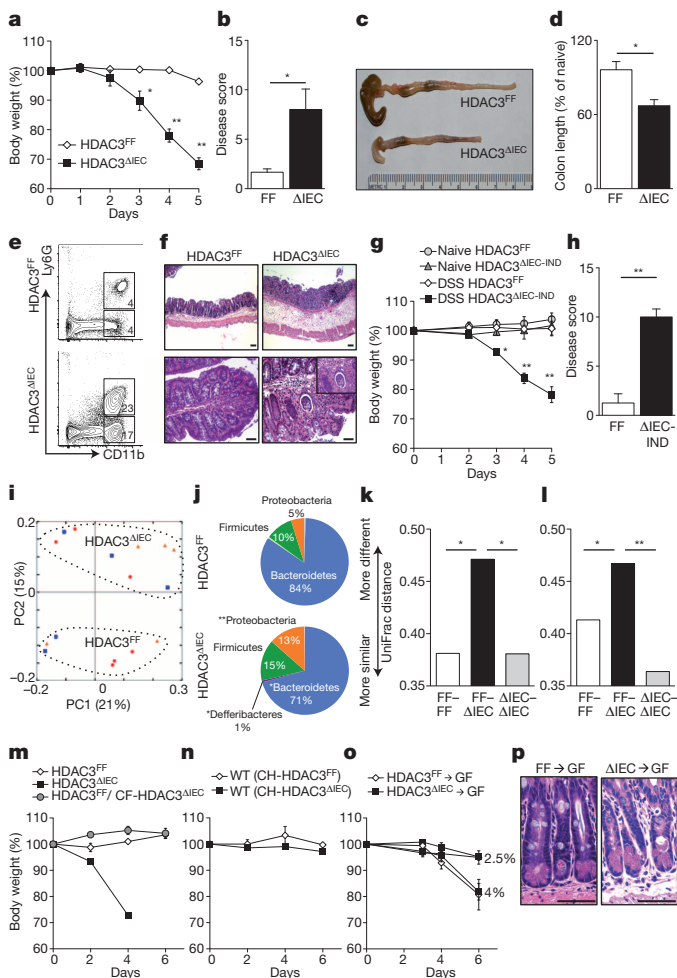


Figure 3 | IEC-intrinsic HDAC3 expression regulates susceptibility to DSS-induced intestinal damage and inflammation. **a**, Changes in body weight after 2.5% DSS. **b–d**, Total disease score (**b**) and colon length (% of naive) (**c**, **d**) on day 5 of DSS. **e**, Frequencies of neutrophils (CD11b⁺Ly6G⁺) and macrophages (CD11b⁺Ly6G[−]) in the colonic lamina propria. **f**, Intestinal sections (caecum, top; colon, bottom; inset: $\times 40$ of crypt abscess). **g**, Changes in body weight for tamoxifen-induced mice. **h**, Total disease score on day 5 of DSS. Data are representative of four independent experiments containing four mice per group. **i**, Comparison of stool bacterial communities at multiple time points (red circles, 6 weeks; blue squares, 8 weeks; orange triangles, 10 weeks). **j**, Phylum comparison of compiled samples from **i**. **k**, **l**, Average UniFrac distance (the distance between bacterial communities based on phylogenetics; see ref. 31) between HDAC3^{FF} and HDAC3^{ΔIEC} mice (FF- Δ IEC) or within genotypes (FF-FF or Δ IEC- Δ IEC) based on 16S rRNA gene sequences from small (**k**) or large (**l**) intestinal luminal samples. $n = 3$ mice per group. **m**, **n**, Changes in body weight of mice cross-fostered (CF) (**m**) or co-housed (CH) (**n**) with HDAC3^{FF} or HDAC3^{ΔIEC} mice before DSS treatment. **o**, Changes in body weight after DSS, and **p**, ileal sections from colonized germ-free (GF) mice. Data are representative of 2–3 independent experiments containing 3 mice per group. Scale bars, 50 μm. Results are shown as mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$.

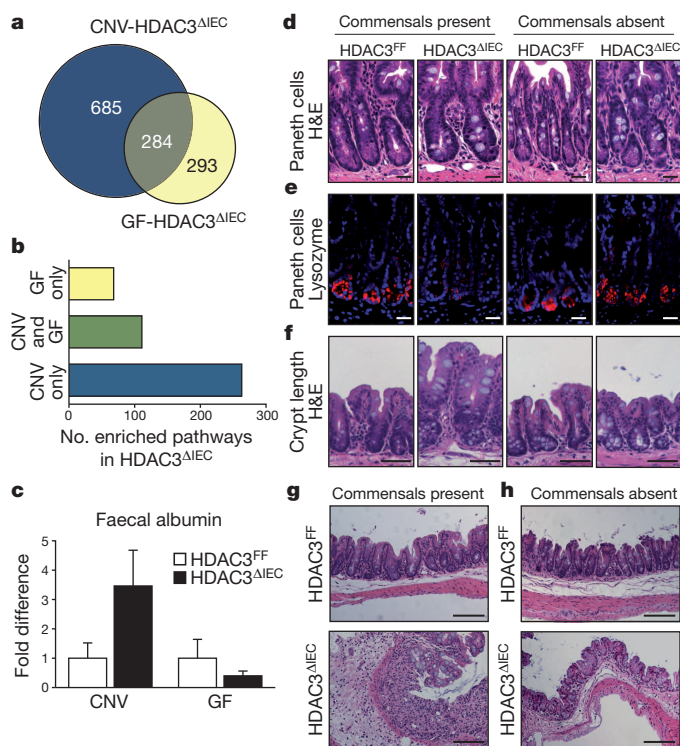


Figure 4 | HDAC3-dependent regulation of intestinal homeostasis depends on integration of commensal-bacteria-derived signals. **a**, Venn diagram showing overlap of differentially expressed genes in EpCAM⁺ IECs isolated from conventionally housed (CNV)-HDAC3^{ΔIEC} mice versus germ-free (GF)-HDAC3^{ΔIEC} mice. Numbers of genes were determined by comparison of HDAC3^{ΔIEC} mice to respective conventionally housed or germ-free HDAC3^{FF} controls. *n* = 3 mice per group. **b**, Number of significantly enriched pathways in HDAC3^{ΔIEC} mice using DAVID pathway analysis (*P* < 0.05). **c**, Albumin measured by ELISA from faecal samples. *n* = 3 mice per group. **d, e**, Haematoxylin and eosin (H&E) (d) and lysozyme (e) stained small intestine. Scale bars, 20 μm. **f**, H&E-stained large intestine. Scale bars, 50 μm. **g, h**, H&E-stained large intestine from HDAC3^{FF} or HDAC3^{ΔIEC} mice in conventionally housed (g) or germ-free (h) conditions after 4 days of 2.5% DSS administration. Scale bars, 100 μm. Data are representative of two independent experiments. Results are shown as mean ± s.e.m.

either the HDAC3^{FF} or the HDAC3^{ΔIEC} microbiota did not exhibit differences in susceptibility to DSS-induced inflammation (Fig. 3o) or Paneth cell homeostasis (Fig. 3p).

These results indicate that the intestinal dysbiosis alone is insufficient to cause the dysregulation in IEC homeostasis or susceptibility to intestinal inflammation that occurs in HDAC3^{ΔIEC} mice. Therefore, we proposed that HDAC3 expression may be required to integrate signals derived from commensal bacteria to regulate intestinal homeostasis. To test this, HDAC3^{ΔIEC} mice were re-derived into germ-free conditions. Genome-wide transcriptional profiling comparing differentially expressed genes of conventionally housed HDAC3^{ΔIEC} mice versus germ-free HDAC3^{ΔIEC} mice revealed that a significant proportion of HDAC3-dependent gene expression relies on the presence of live commensal bacteria (Fig. 4a). Furthermore, alterations in most HDAC3-dependent transcriptional pathways, including those involved in antimicrobial defence, occurred only in HDAC3^{ΔIEC} mice housed under conventional conditions (Fig. 4b and Extended Data Fig. 8), indicating that a dominant role for HDAC3 in regulation of IEC-intrinsic transcriptional pathways requires commensal-bacteria-derived signals. Consistent with this, whereas faecal albumin levels (Fig. 4c), Paneth cell homeostasis (Fig. 4d, e) and crypt length (Fig. 4f) were dysregulated in conventionally housed HDAC3^{ΔIEC} mice compared to conventionally housed HDAC3^{FF} mice, these differences were significantly abrogated between germ-free HDAC3^{FF} versus germ-free

HDAC3^{ΔIEC} mice (Fig. 4c–f). In addition, whereas conventionally housed HDAC3^{ΔIEC} mice were more susceptible to DSS-induced intestinal inflammation compared to conventionally housed HDAC3^{FF} mice (Fig. 4g), minimal differences in intestinal inflammation were observed between germ-free HDAC3^{FF} versus germ-free HDAC3^{ΔIEC} mice (Fig. 4h).

Collectively, these data indicate that expression of HDAC3 in IECs is critical for the coordinated expression of networks of genes that regulate IEC function and tissue homeostasis in the presence of commensal bacteria. Notably, under germ-free conditions, IEC-intrinsic HDAC3 expression was dispensable for regulation of epithelial barrier integrity and intestinal homeostasis, indicating that HDAC3 integrates commensal-bacteria-derived signals to maintain normal host–commensal relationships (Extended Data Fig. 9). Therefore, in addition to established pathways of immune recognition of commensal bacteria via germline encoded pattern recognition receptors, IEC-intrinsic expression of HDAC3 may have an evolutionarily conserved role in regulating host–commensal bacteria relationships. In this context, IEC-intrinsic HDAC3 may influence susceptibility to multiple systemic chronic inflammatory diseases that are influenced by both host genetic and microbe-derived factors.

METHODS SUMMARY

HDAC3^{ΔIEC} and HDAC3^{ΔIEC-IND} mice were generated by breeding HDAC3^{FF} mice²⁴ to C57BL/6 mice expressing Cre-recombinase or tamoxifen-dependent Cre recombinase under the control of the villin promoter. Deletion of HDAC3 in HDAC3^{ΔIEC-IND} mice was induced by intraperitoneal injection of tamoxifen. Germ-free HDAC3^{ΔIEC} mice were re-derived at the University of Pennsylvania Gnotobiotic Mouse Facility. Albumin levels in faecal homogenates were quantified by ELISA. DSS was added to drinking water at 2.5% weight/volume for 5 days. Human intestinal tissue was obtained from the University of Pennsylvania IBD Immunology Initiative. IECs were isolated by shaking intestinal tissue in 1 mM EDTA/1 mM dithiothreitol (DTT) and 5% FCS at 37 °C. For microarray analyses, three biological replicates of EpCAM⁺ IECs were sorted for each genotype and condition. ChIP DNA libraries from three biological replicates for each genotype were used for sequencing. DNA from stool and intestinal contents was amplified using barcoded V1–V2 region primers targeting bacterial 16S rRNA gene, sequenced using 454/Roche Titanium technology, analysed using QIIME, and assessed by multiple parameters (Extended Data Fig. 10).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 July 2012; accepted 19 September 2013.

Published online 3 November 2013.

1. Strober, W., Fuss, I. & Mannon, P. The fundamental basis of inflammatory bowel disease. *J. Clin. Invest.* **117**, 514–521 (2007).
2. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474**, 327–336 (2011).
3. Kaser, A., Zeissig, S. & Blumberg, R. S. Inflammatory bowel disease. *Annu. Rev. Immunol.* **28**, 573–621 (2010).
4. Renz, H. *et al.* Gene-environment interactions in chronic inflammatory disease. *Nature Immunol.* **12**, 273–277 (2011).
5. Slomko, H., Heo, H. J. & Einstein, F. H. Minireview: Epigenetics of obesity and diabetes in humans. *Endocrinology* **153**, 1025–1030 (2012).
6. Mukherjee, A. B. & Zhang, Z. Allergic asthma: influence of genetic and environmental factors. *J. Biol. Chem.* **286**, 32883–32889 (2011).
7. Denson, L. A. *et al.* Challenges in IBD research: update on progress and prioritization of the CCFA's research agenda. *Inflamm. Bowel Dis.* **19**, 677–682 (2013).
8. Ivanov, I. I. & Honda, K. Intestinal commensal microbes as immune modulators. *Cell Host Microbe* **12**, 496–508 (2012).
9. Cadwell, K. *et al.* Virus-plus-susceptibility gene interaction determines Crohn's disease gene *Atg16L1* phenotypes in intestine. *Cell* **141**, 1135–1145 (2010).
10. Gallo, R. L. & Hooper, L. V. Epithelial antimicrobial defence of the skin and intestine. *Nature Rev. Immunol.* **12**, 503–516 (2012).
11. Artis, D. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nature Rev. Immunol.* **8**, 411–420 (2008).
12. Donohoe, D. R. & Bultman, S. J. Metabolite epigenetics: interrelationships between energy metabolism and epigenetic control of gene expression. *J. Cell. Physiol.* **227**, 3169–3177 (2012).
13. Perissi, V. & Rosenfeld, M. G. Controlling nuclear receptors: the circular logic of cofactor cycles. *Nature Rev. Mol. Cell Biol.* **6**, 542–554 (2005).

14. Haberland, M., Montgomery, R. L. & Olson, E. N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nature Rev. Genet.* **10**, 32–42 (2009).
15. Kim, G. W., Gocevski, G., Wu, C. J. & Yang, X. J. Dietary, metabolic, and potentially environmental modulation of the lysine acetylation machinery. *Int. J. Cell Biol.* **2010**, 632739 (2010).
16. Dashwood, R. H. & Ho, E. Dietary histone deacetylase inhibitors: from cells to mice to man. *Semin. Cancer Biol.* **17**, 363–369 (2007).
17. Chen, X. *et al.* Requirement for the histone deacetylase Hdac3 for the inflammatory gene expression program in macrophages. *Proc. Natl Acad. Sci. USA* **109**, E2865–E2874 (2012).
18. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834–840 (2009).
19. You, S. H. *et al.* Nuclear receptor co-repressors are required for the histone-deacetylase activity of HDAC3 *in vivo*. *Nature Struct. Mol. Biol.* **20**, 182–187 (2013).
20. Alenghat, T. *et al.* Nuclear receptor corepressor and histone deacetylase 3 govern circadian metabolic physiology. *Nature* **456**, 997–1000 (2008).
21. Fischle, W. *et al.* Enzymatic activity associated with class II HDACs is dependent on a multiprotein complex containing HDAC3 and SMRT/N-CoR. *Mol. Cell* **9**, 45–57 (2002).
22. Montgomery, R. L. *et al.* Maintenance of cardiac energy metabolism by histone deacetylase 3 in mice. *J. Clin. Invest.* **118**, 3588–3597 (2008).
23. Feng, D. *et al.* A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. *Science* **331**, 1315–1319 (2011).
24. Mullican, S. E. *et al.* Histone deacetylase 3 is an epigenomic brake in macrophage alternative activation. *Genes Dev.* **25**, 2480–2488 (2011).
25. Pull, S. L., Doherty, J. M., Mills, J. C., Gordon, J. I. & Stappenbeck, T. S. Activated macrophages are an adaptive element of the colonic epithelial progenitor niche necessary for regenerative responses to injury. *Proc. Natl Acad. Sci. USA* **102**, 99–104 (2005).
26. Diehl, G. E. *et al.* Microbiota restricts trafficking of bacteria to mesenteric lymph nodes by CX₃CR1^{hi} cells. *Nature* **494**, 116–120 (2013).
27. Garrett, W. S. *et al.* Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host Microbe* **8**, 292–300 (2010).
28. Elinav, E. *et al.* NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145**, 745–757 (2011).
29. Devkota, S. *et al.* Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in *Il10*^{-/-} mice. *Nature* **487**, 104–108 (2012).
30. Raetz, M. *et al.* Parasite-induced T_H1 cells and intestinal dysbiosis cooperate in IFN- γ -dependent elimination of Paneth cells. *Nature Immunol.* **14**, 136–142 (2013).
31. Lozupone, C., Hamady, M. & Knight, R. UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, 371 (2006).

Acknowledgements We thank members of the Artis laboratory for discussions and critical reading of the manuscript. This research is supported by the National Institutes of Health (AI061570, AI095608, AI087990, AI074878, AI095466, AI106697, AI102942 and AI097333 to D.A.; DK043806 to M.A.L.; T32-RR007063, K08-DK093784 to T.A.; DP5OD012116 to G.F.S.; F31-GM082187 to S.A.S.; K08-DK084347 to M.B.; R21-AI105346 to I.E.B.), the Crohns and Colitis Foundation of America (T.A. and D.A.), the Burroughs Wellcome Fund Investigator in Pathogenesis of Infectious Disease Award (D.A.), and the Irvington Institute Postdoctoral Fellowship of the Cancer Research Institute (L.C.O.). We also thank S. Lukovac and R. Aoki for technical assistance, the University of Pennsylvania Matthew J. Ryan Veterinary Hospital Pathology Lab, Center for AIDS Research, the Penn Microarray Facility, the NIH/NIDDK Center for Molecular Studies in Digestive and Liver Diseases (P30-DK050306) and its core facilities (Molecular Pathology and Imaging; Molecular Biology; Cell Culture; Transgenic and Chimeric Mouse), the Functional Genomics Core of the Penn Diabetes Research Center (DRC) (P30-DK19525), the Pathology Core at the Stokes Institute, the Electron Microscopy Resource Laboratory, the Penn IBD Immunology Initiative (I³), and the Mucosal Immunology Studies Team (MIST) of the NIAID for sharing expertise and resources. The authors would also like to thank the Abramson Cancer Center Flow Cytometry and Cell Sorting Resource Laboratory for technical advice and support. The ACC Flow Cytometry and Cell Sorting Shared Resource is partially supported by NCI Comprehensive Cancer Center Support Grant (2-P30 CA016520). Some human tissue samples were provided by the Cooperative Human Tissue Network (funded by the National Cancer Institute).

Author Contributions T.A., L.C.O., S.A.S., D.K., M.W.-D., A.S., P.R.G., K.L.J. and G.F.S. designed and performed the research, T.A. and D.K. re-derived the germ-free mice, C.G.K.Z. performed analyses of microarray data, I.C. and K.-J.W. performed analyses of ChIP-seq data, S.G., R.S. and F.D.B. performed and analysed 454 pyrosequencing, T.B.H. and M.B. provided human tissues, S.E.M., I.E.B., G.F.S., F.D.B. and M.A.L. provided mouse strains, advice or technical expertise, and T.A. and D.A. analysed the data and wrote the manuscript.

Author Information Detailed microarray data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE50190. ChIP-seq data have been deposited in GEO under accession number GSE50453. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A. (dartis@mail.med.upenn.edu).

METHODS

Mice. Previously described HDAC3^{FF} mice²⁴ were bred to C57BL/6 mice expressing Cre-recombinase³² or tamoxifen-dependent Cre recombinase³³ under the control of the villin promoter (Jackson Laboratory) to generate HDAC3^{AIEC} and HDAC3^{AIEC-IND} mice, respectively. Genotypes were determined by PCR. Deletion of HDAC3 in HDAC3^{AIEC-IND} mice was induced by intraperitoneal injection of 1 mg of tamoxifen (Sigma) once per day for 5 days after weaning to adult mice that ranged from 8–10 weeks old. For Paneth cell examination in HDAC3^{AIEC-IND} mice, tamoxifen was administered for three 5-day intervals during a 30-day period. HDAC3^{ΔlysM} mice have been described previously²⁴ and express Cre recombinase under the control of the lysozyme M promoter. Germ-free HDAC3^{AIEC} mice were re-derived at the University of Pennsylvania Gnotobiotic Mouse Facility. For co-housing experiments, 4-week-old wild-type and HDAC3^{FF} or HDAC3^{AIEC} mice were co-housed for 4 weeks. For cross-fostering experiments, newborn HDAC3^{FF} mice were transferred to a litter with a HDAC3^{AIEC} dam and HDAC3^{AIEC} pups at birth. Colonization of 8-week-old wild-type germ-free mice was performed with caecal contents collected from either HDAC3^{FF} or HDAC3^{AIEC} mice. All mice were used at 7–12 weeks old, and age- and gender-matched mice were used for all experiments. Animals were housed up to 5 per cage in a ventilated isolator cage system in a 12 h light/dark cycle, with free access to water and chow. Animals requiring medical attention were provided with appropriate veterinary care by a licensed veterinarian and were excluded from the experiments described. No other exclusion criteria existed. All experiments were done according to the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee. Germ-free mice were maintained in plastic isolator units, were fed autoclaved feed and water and were routinely monitored to ensure the absence of microbial contamination.

Human intestinal samples, IEC harvest, RNA, ChIP, real-time PCR and western blotting. De-identified human intestinal tissue from the terminal ileum and rectum were obtained from the University of Pennsylvania IBD Immunology Initiative (IRB 814428). All IBD participants had a confirmed diagnosis of Crohn's disease or ulcerative colitis by a gastroenterologist and had provided informed consent. Patients younger than 18 years of age were excluded. Control samples were obtained from consenting men and women older than 18 years of age with no history of a diagnosis of inflammatory bowel disease, microscopic colitis, or ischaemic colitis. Human IECs were purified from biopsy samples by incubating tissue in 1 mM EDTA/1 mM DTT and 5% FCS at 37 °C for 20–30 min and vortexing every 5 min. Human IECs were 99.3% negative for CD45⁺ cells. IECs were isolated from murine samples by shaking intestinal tissue in 1 mM EDTA/1 mM DTT and 5% FCS at 37 °C for 10 min, resulting in 80–90% EpCAM⁺ IEC purity. RNA was isolated from cells using the RNeasy kit (Qiagen) then subjected to reverse transcription with Superscript reverse transcriptase (Invitrogen). ChIP was performed as described previously^{20,24} with few modifications. Briefly, cells were fixed in 1% PFA for 10 min and quenched with glycine. Total cell extracts were sonicated using a Bioruptor (Diagenode) and appropriate sonication was confirmed using an Agilent bioanalyzer. Extracts were immunoprecipitated with rabbit anti-H3K9ac (Millipore; 06-942). Real-time PCR was performed using SYBR green chemistry (Applied Biosystems), commercially available primer sets (Qiagen) or custom made primer pairs (Invitrogen). Reactions were run on a real-time PCR system (ABI7500; Applied Biosystems) and data were analysed with a threshold set in the linear range of amplification and processed based on a standard curve of serial tenfold dilutions for each primer set. Samples were normalized to an unaffected endogenous control gene and plotted as mean fold difference (\pm s.e.m.) relative to control mice. For western blot analysis, IECs were lysed in a modified RIPA buffer and lysates were subjected to immunoblot analysis. Blots were probed with rabbit anti-HDAC3 (Santa Cruz) and mouse anti-actin (Cell Signaling).

Flow cytometry. To isolate lamina propria immune cells, IEC and intraepithelial lymphocyte layers were first stripped by shaking sections of large intestine in 5 mM EDTA/1 mM DTT. Remaining tissue was then digested with collagenase (0.5 mg ml⁻¹) to obtain single cell suspensions. For flow cytometry, cells were stained with a combination of the following fluorescence-conjugated monoclonal antibodies: phycoerythrin (PE)–Texas red conjugated anti-CD11b (Invitrogen), PE-Cy7 conjugated anti-F4/80 (eBioscience), Alexa Fluor 700 conjugated anti-Ly6G (Biolegend), allophycocyanin (APC) conjugated anti-CD3 (eBioscience), PE-Cy5 conjugated anti-CD19 (eBioscience), APC-Cy7 conjugated anti-CD11c (eBioscience), eFluor650NC conjugated anti-CD45 (eBioscience), PerCP-Cy5.5 conjugated anti-CD4 (eBioscience). Dead cells were excluded from analysis through the use of a LIVE/DEAD Fixable Aqua Dead Cell Stain kit (Invitrogen). Samples were acquired on an LSR II (BD Biosciences) and were analysed with FlowJo software (v9.2; TreeStar).

Microarray analysis. EpCAM⁺ IECs were sorted as DAPI⁻, Lin⁻ (CD45, CD4, CD8, CD11b, CD19), EpCAM⁺ from large intestinal IEC preparations using a BD Aria II with a 100 μ m nozzle. EpCAM staining was performed using APC conjugated

anti-EpCAM (eBioscience; G8.8). Three biological replicates were collected for each genotype and condition, each containing $1.0\text{--}1.7 \times 10^5$ cells sorted to a purity of 99%. Total RNA was prepared using TRIzol (Invitrogen) and analysed by the Microarray Core at the University of Pennsylvania. cDNA was amplified using NuGen WT Ovation Pico kit and hybridized to an Affymetrix GeneChip (Mouse Gene 1.0ST). Affymetrix Power Tools software was used for processing and quantile normalization of fluorescence hybridization signals. Transcripts were log₂-normalized and average values were obtained for analysis of expression. Detailed microarray data have been deposited in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE50190. Genes with greater than 1.5-fold change were analysed within DAVID (Database for Annotation, Visualization and Integrated Discovery). The R package ComBat was implemented to eliminate non-biological experimental artefacts³⁴ and the Broad Institute GSEA software was used for gene-set enrichment analysis as described³⁵.

ChIP-seq. ChIP DNA libraries from three biological replicates for each genotype were prepared for sequencing according to the multiplex amplification protocol from Illumina. Sequencing was performed by the The Functional Genomics Core of the Penn Diabetes Research Center. Sequence reads of 50 base pairs (bp) were obtained using the Solexa Analysis Pipeline. ChIP-seq reads for H3K9ac were mapped to the mouse genome (mm9) using bowtie by allowing up to two mismatches³⁶. Reads were normalized to reads per kilobase per 10 million mapped reads and clonal reads were removed. Average H3K9ac signals for each gene at the TSS \pm 2 kb region were calculated from three independent biological replicates and *P* values were determined using a paired *t*-test. ChIP-seq data have been deposited in GEO and are accessible through GEO series accession number GSE50453.

Histology, immunohistochemistry and immunofluorescence. Sections of intestine were fixed in 4% paraformaldehyde, paraffin embedded, sectioned and stained with haematoxylin and eosin (H&E), periodic acid-Schiff (PAS)/Alcian blue, anti-HDAC3 (Santa Cruz), anti-Ki67 (AbCam), or anti-caspase 3 (R&D systems) for immunohistochemistry, or anti-lysozyme (Santa Cruz) and DAPI for immunofluorescence. Pathology was scored based on oedema (1–4) and inflammation (1–4). For EM, intestinal tissues were fixed with 2.5% glutaraldehyde, 2.0% paraformaldehyde in 0.1 M sodium cacodylate buffer, pH 7.4. After buffer washes, the samples were post-fixed in 2.0% osmium tetroxide for 1 h at room temperature, and rinsed in dH₂O before en bloc staining with 2% uranyl acetate. After dehydration through a graded ethanol series, the tissue was infiltrated and embedded in EMbed-812 (Electron Microscopy Sciences). Thin sections were stained with uranyl acetate and lead citrate and examined with a JEOL 1010 electron microscope fitted with a Hamamatsu digital camera and AMT Advantage image capture software.

In vivo intestinal barrier function assays. Mice were fasted overnight and FITC-dextran (0.6 mg g⁻¹; Sigma) diluted in PBS was gavaged the following day. Fluorescence intensity of plasma samples was measured (excitation 485 nm/emission, 535 nm) 4 h after gavage. For faecal albumin assays, faecal pellets were weighed and homogenized in diluent (PBS, 1% BSA, 0.05% Tween 20). Albumin levels in faecal homogenates were quantified by ELISA according to the manufacturer's protocol (Bethyl Laboratories). LPS levels in mesenteric lymph node homogenates were assayed via the Limulus amoebocyte lysate (LAL) test according to the manufacturer's protocol (Lonza). Albumin and LPS levels were normalized to faecal or tissue weight and presented as fold difference relative to HDAC3^{FF} mice.

Murine colitis model. DSS (MP Biomedicals; relative molecular mass 36,000–50,000) was added to drinking water at 2.5% weight/volume for 5 days. Disease was scored as follows: (1) weight loss (no change = 0; <5% = 1; 6–10% = 2; 11–20% = 3; > 20% = 4); (2) faeces (normal = 0; pasty, semiformal = 2; liquid, sticky, or unable to defecate after 5 min = 4); (3) blood (no blood = 0; visible blood in rectum = 1; visible blood on fur = 2); and (4) general appearance (normal = 0; piloerection = 1; lethargy and piloerection = 2; motionless = 4).

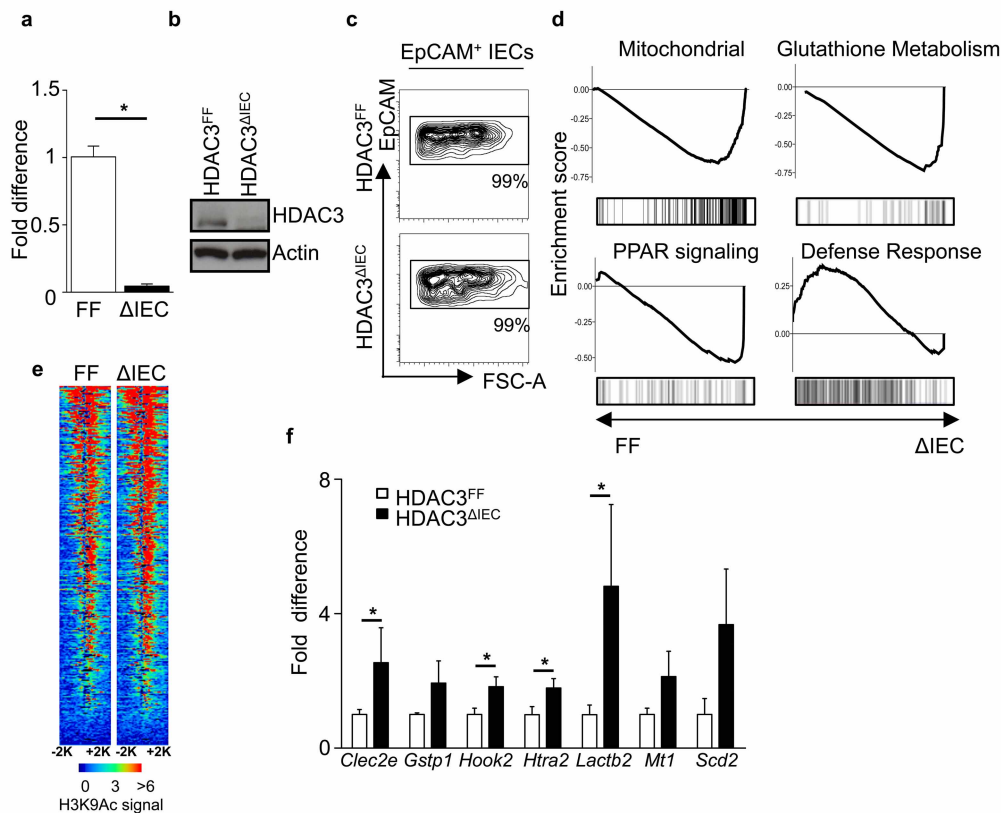
16S rRNA gene pyrosequencing. Stool and intestinal contents total DNA was extracted using the QIAamp DNA Stool Mini kit (Qiagen). DNA samples were amplified using barcoded V1–V2 region primers targeting bacterial 16S rRNA gene and sequenced using 454/Roche Titanium technology. Sequence analysis was carried out using the QIIME pipeline³⁷ with default settings and data was submitted to the Sequence Read Archive (SRA) and are accessible through accession number SRP029234. Taxonomic assignments were carried out using RDP. Community structure comparisons were carried out using UniFrac^{31,38} and principal coordinate analysis. For distance analysis in UniFrac, *P* values were determined using label permutation as implemented in the QIIME package. Pyrosequencing parameters including number of reads, alpha diversity, rarefaction curves, additional PCA plots, and clustering dendrograms are displayed in Extended Data Fig. 10.

Oral *L. monocytogenes* infection and bactericidal assays. Mice were infected orally with streptomycin-resistant *L. monocytogenes* with 3×10^8 colony forming units (c.f.u.) and weighed daily. Seventy-two hours after infection, mesenteric lymph nodes were homogenized in PBS, and serial dilutions of the homogenates were plated on LB plates containing 100 μ g ml⁻¹ streptomycin, incubated at 37 °C

and c.f.u. were counted. For bactericidal assays, small intestinal crypts were isolated, stimulated with 10 μ m carbamyl choline (CCh) and assayed against *Salmonella enterica* serovar Typhimurium as described previously³⁹.

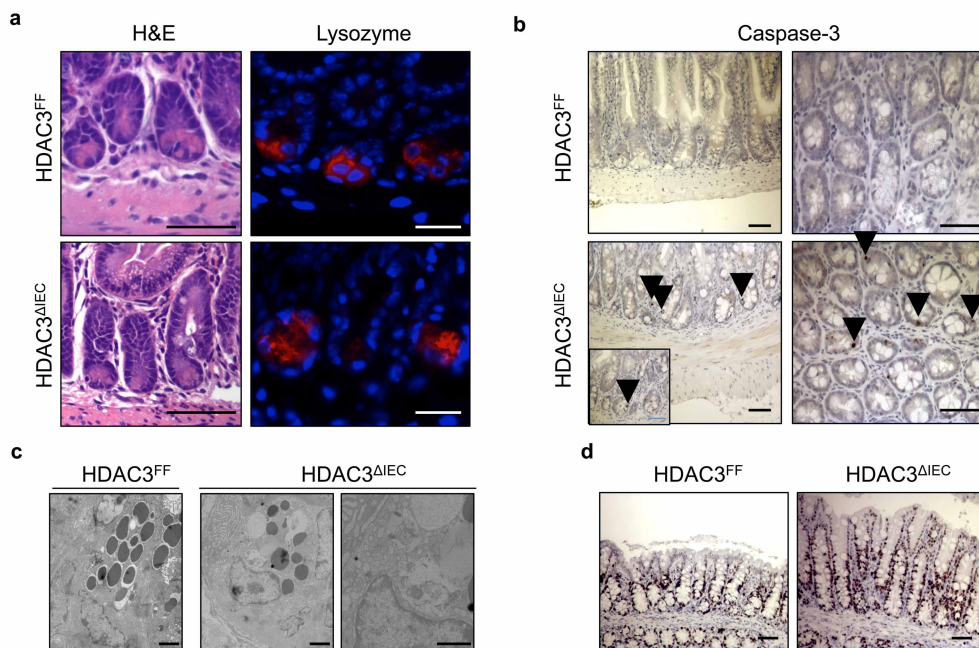
Statistics. Results are shown as mean \pm s.e.m. To determine group sizes necessary for adequate statistical power, power analysis was performed using preliminary data sets. Mice of the indicated genotypes were assigned at random to groups. Mouse studies were not performed in a blinded fashion. All inclusion/exclusion criteria were pre-established. Statistical significance was determined with the *t*-test or Mann–Whitney *U*-test. All data meet the assumptions of the statistical tests used. Within each group there is an estimate of variation, and the variance between groups is similar. For each statistical analysis, appropriate tests were selected based on whether the data was normally distributed and whether multiple comparisons were made. Results were considered significant at **P* \leq 0.05; ***P* \leq 0.01. Statistical analyses were performed using Prism version 5.0a (GraphPad Software).

32. Madison, B. B. *et al.* Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277**, 33275–33283 (2002).
33. El Marjou, F. *et al.* Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis* **39**, 186–193 (2004).
34. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
36. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
37. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
38. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
39. Ayabe, T. *et al.* Secretion of microbicidal α -defensins by intestinal Paneth cells in response to bacteria. *Nature Immunol.* **1**, 113–118 (2000).



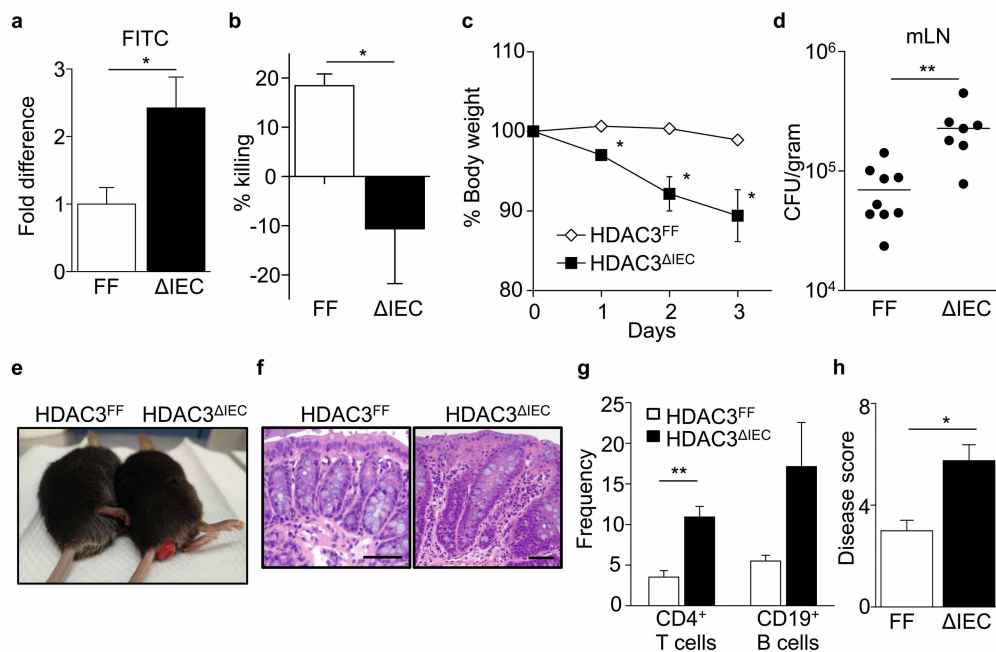
Extended Data Figure 1 | IECs from HDAC3^{ΔIEC} mice demonstrate alterations in gene expression coupled with increased histone acetylation. **a, b**, HDAC3 expression in IECs from HDAC3^{FF} or HDAC3^{ΔIEC} mice by real-time PCR (**a**) and western analysis (**b**). **c**, Purity of sort-purified EpCAM⁺ IECs. **d**, Gene-set enrichment analysis (GSEA) comparing IECs from HDAC3^{FF} and HDAC3^{ΔIEC} mice to published data enrichment sets obtained from the Molecular Signatures Database. **e**, Heat map of H3K9Ac signal in IECs from HDAC3^{FF} and HDAC3^{ΔIEC} mice at genes that are upregulated in IECs of

HDAC3^{ΔIEC} mice. Each row represents a single gene sorted by the peak heights in the HDAC3^{ΔIEC} mice. H3K9Ac signals were normalized to reads per kilobase per 10 million mapped reads. **f**, ChIP-qPCR comparing H3K9Ac levels in IECs from HDAC3^{ΔIEC} mice versus HDAC3^{FF} mice at promoter regions of select upregulated genes. Data are presented as fold difference relative to control HDAC3^{FF} IECs. *n* = 3 mice per group. Results are shown as mean ± s.e.m. **P* < 0.05.



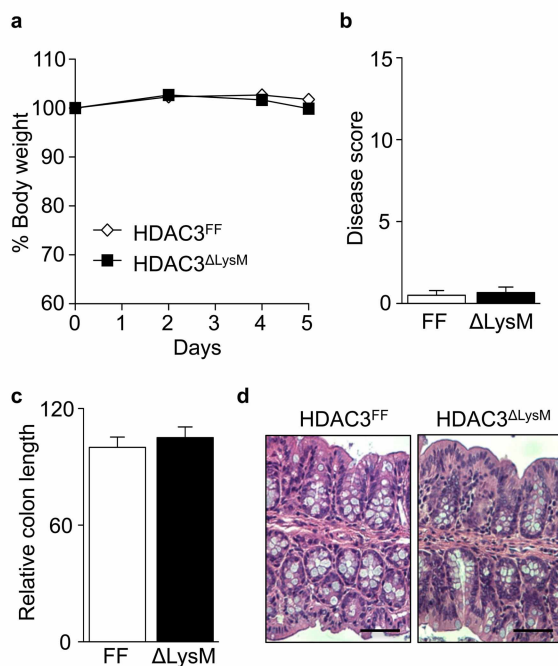
Extended Data Figure 2 | HDAC3^{ΔIEC} mice exhibit altered IEC homeostasis. **a**, Representative haematoxylin and eosin (H&E; left) and lysozyme (right) stained sections of small intestine from 18-day-old HDAC3^{FF} and HDAC3^{ΔIEC} mice. Scale bars, 50 μm. **b**, Immunohistochemistry for active caspase-3 in small intestinal crypts from HDAC3^{FF} and HDAC3^{ΔIEC} mice.

Arrows indicate positive nuclear staining. Scale bars, 50 μm. **c**, Electron micrograph of littermate HDAC3^{FF} and HDAC3^{ΔIEC} Paneth cells. Scale bars, 2 μm. **d**, Immunohistochemistry for Ki67 in colonic crypts from HDAC3^{FF} and HDAC3^{ΔIEC} mice. Scale bars, 50 μm.



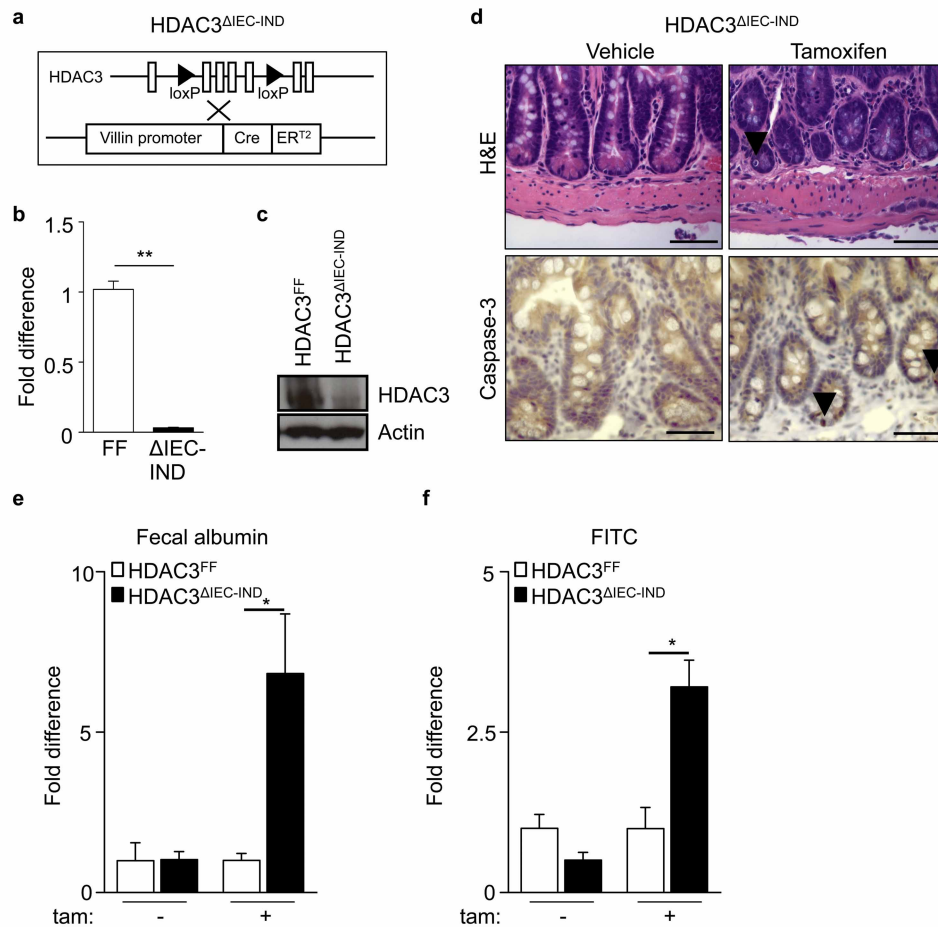
Extended Data Figure 3 | HDAC3 ^{Δ IEC} mice demonstrate impaired intestinal barrier function, spontaneous intestinal inflammation and defective antibacterial defences. **a**, FITC levels in plasma assessed 4 h after oral gavage with FITC-dextran (0.6 mg g^{-1}) of HDAC3^{FF} ($n = 3$) and HDAC3 ^{Δ IEC} ($n = 5$) mice and presented as fold difference relative to HDAC3^{FF} mice. **b**, Bactericidal activity against *Salmonella enterica* serovar Typhimurium of supernatants from carbamyl choline (CCh)-stimulated small intestinal crypts. Data are presented as percentage killing compared to unstimulated crypts. $n = 4$ mice per group. **c**, Daily changes in body weight after oral infection with

L. monocytogenes. **d**, Colony forming units (c.f.u.) of *L. monocytogenes* grown on LB plates containing streptomycin from mesenteric lymph nodes (mLN) 72 h after infection. HDAC3^{FF} ($n = 9$) and HDAC3 ^{Δ IEC} ($n = 7$). **e**, Rectal prolapse in a 4-month-old HDAC3 ^{Δ IEC} mouse. **f–h**, Representative haematoxylin and eosin stained section of colons (**f**), quantification of CD4⁺ and CD19⁺ cells (gated live, CD45⁺) in lamina propria (**g**), and disease score (**h**) from mice in **e**. Scale bars, 50 μm . Data depicted are from two pooled experiments. Results are shown as mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$.



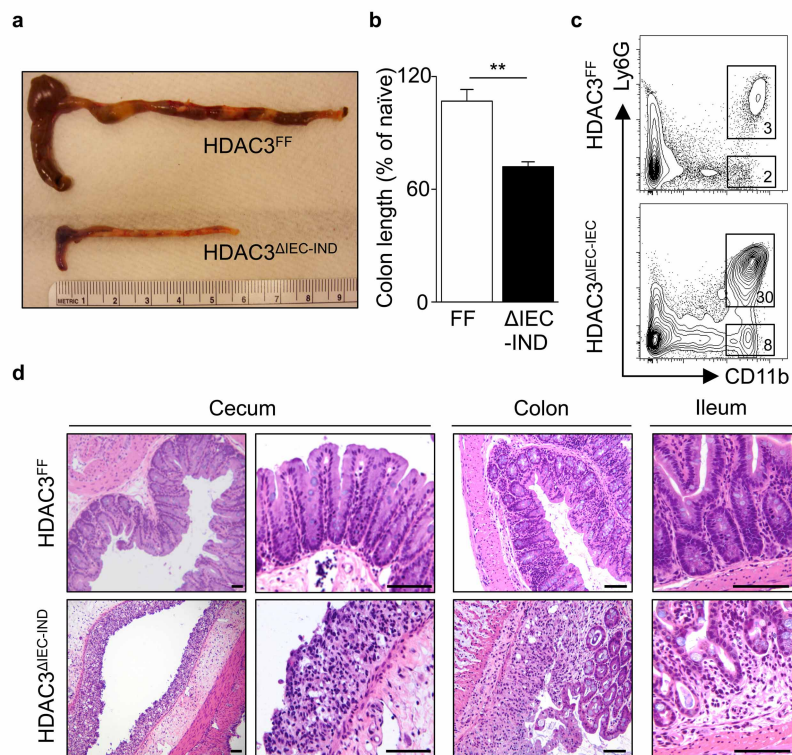
Extended Data Figure 4 | HDAC3 ^{Δ LysM} mice do not demonstrate increased sensitivity to DSS-induced intestinal damage and inflammation. **a–d**, Daily changes in body weight (**a**), disease score (**b**), colon length (**c**) and representative haematoxylin and eosin stained large intestine sections (**d**) of

2.5% DSS-treated HDAC3^{FF} and HDAC3 ^{Δ LysM} mice. Scale bars, 50 μm . $n = 4$ mice per group. Data are representative of two independent experiments. Results are shown as mean \pm s.e.m.



Extended Data Figure 5 | HDAC3 Δ IEC-IND mice exhibit Paneth cell loss and impaired barrier function after tamoxifen-induced deletion of HDAC3 in IECs. **a**, HDAC3 Δ IEC-IND mice contain the floxed HDAC3 gene and a tamoxifen-dependent Cre recombinase (Cre-ERT²) controlled by the villin promoter. **b**, **c**, HDAC3 expression in IECs from HDAC3^{FF} and HDAC3 Δ IEC-IND mice by real-time PCR (**b**) and western analysis (**c**) after tamoxifen treatment. **d**, Representative haematoxylin and eosin (top) and active caspase-3 (bottom) stained sections of small intestine of HDAC3 Δ IEC-IND

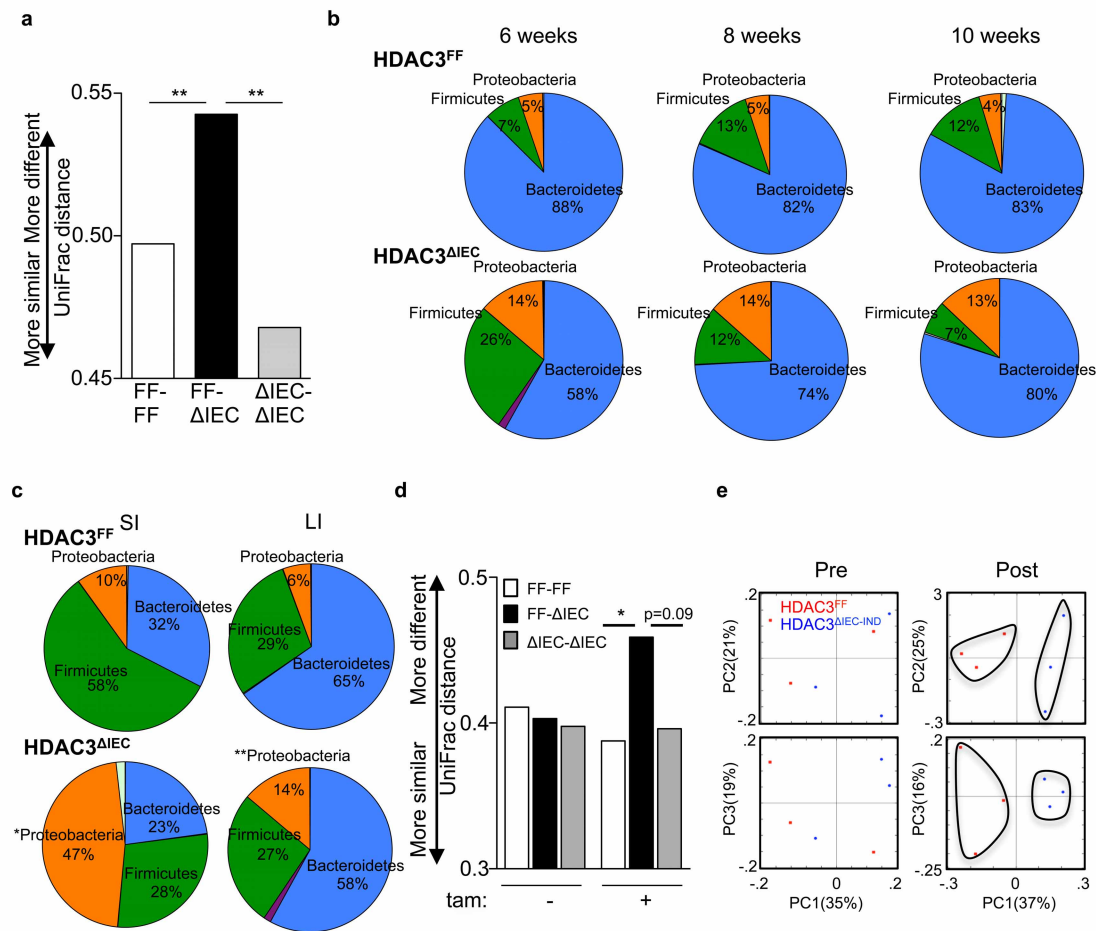
mice treated with either vehicle or tamoxifen for three 5-day periods over 30 days. Arrows indicate dead cell (top) and positive nuclear staining (bottom). Scale bars, 50 μ m. **e**, Albumin measured by ELISA from faecal samples collected from the same mice before tamoxifen-induced HDAC3 deletion (–) and after tamoxifen-induced deletion (+). **f**, FITC levels in plasma assessed 4 h after oral gavage. HDAC3^{FF} ($n = 3$), HDAC3 Δ IEC-IND ($n = 8$). Data are representative of two independent experiments. Results are shown as mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$.



Extended Data Figure 6 | HDAC3^{ΔIEC-IND} mice exhibit enhanced susceptibility to DSS-induced intestinal damage and inflammation.

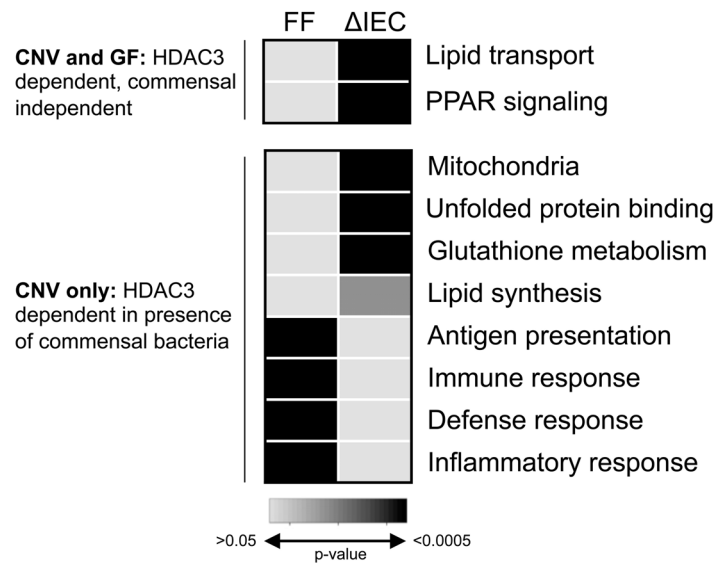
a, b, Representative large intestine (**a**) and colon length (% naïve) (**b**) after 5 days of 2.5% DSS. **c**, Frequencies of neutrophils (CD11b⁺Ly6G⁺) and macrophages (CD11b⁺Ly6G⁻) in the colonic lamina propria.

d, Representative haematoxylin and eosin stained intestine sections of HDAC3^{FF} and HDAC3^{ΔIEC-IND} mice. Scale bars, 50 μ m. $n = 4$ mice per group. Data are representative of four independent experiments. Results are shown as mean \pm s.e.m. $**P < 0.01$.

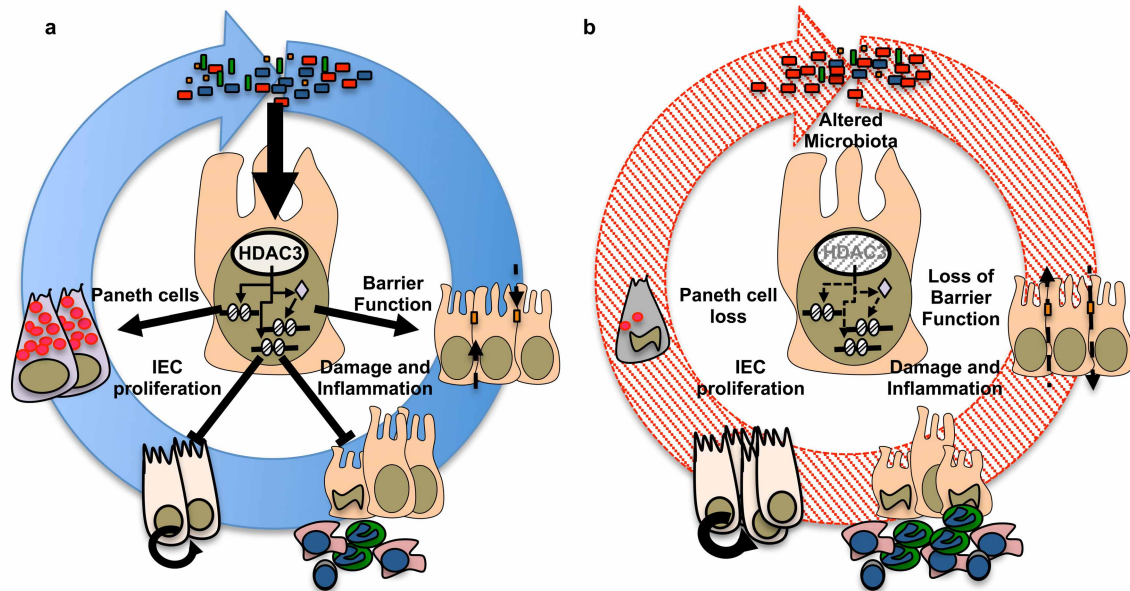


Extended Data Figure 7 | Inhibition of IEC-intrinsic HDAC3 results in temporal and spatial alterations in the diversity of intestinal commensal bacteria. **a**, Average UniFrac distance within each genotype (HDAC3^{FF} mice and HDAC3^{ΔIEC} mice), or between HDAC3^{FF} and HDAC3^{ΔIEC} mice based on 16S rRNA gene sequences determined from stool bacterial communities collected three times over a 4-week period from adult HDAC3^{FF} and HDAC3^{ΔIEC} mice. **b**, Phylum-level comparison of stool bacterial communities

at each time point. **c**, Phylum-level comparison of bacterial communities in contents from small (SI) or large intestine (LI). **d**, Average UniFrac distance between HDAC3^{FF} mice and HDAC3^{ΔIEC-IND} mice, or within HDAC3^{FF} and HDAC3^{ΔIEC-IND} groups based on 16S rRNA gene sequences determined from stool bacterial communities collected before tamoxifen induction (Pre) and 15 days after 5 days of tamoxifen administration (Post). **e**, Principal coordinate analysis of samples in **d**. *n* = 3 mice per group. **P* < 0.05, ***P* < 0.01.

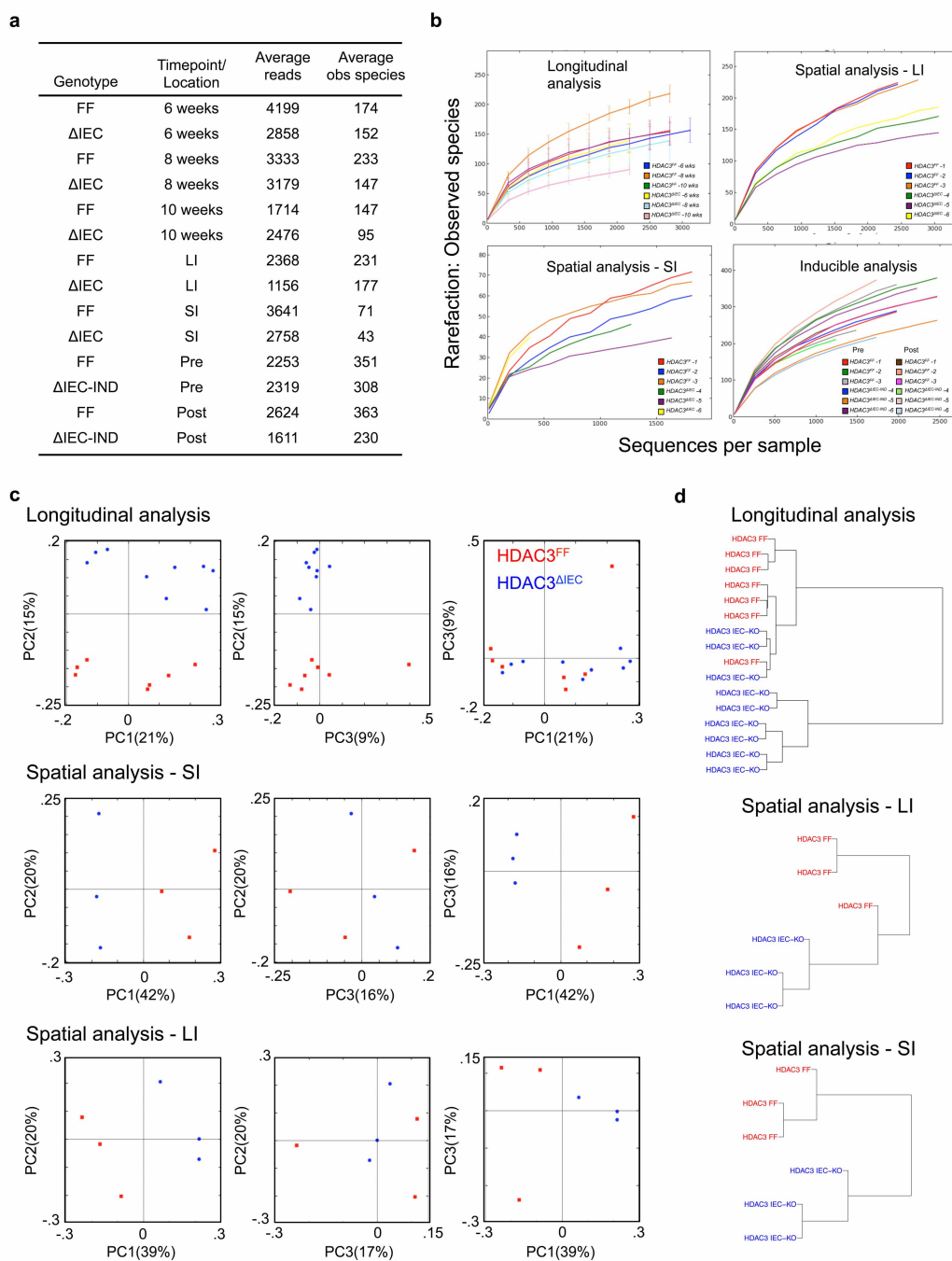


Extended Data Figure 8 | HDAC3-dependent regulation involves integration of commensal-bacteria-derived signals. Functional classification of enriched pathways by DAVID pathway analysis using genes represented in Fig. 4a.



Extended Data Figure 9 | Epithelial HDAC3 integrates commensal-bacteria-derived signals to establish commensalism and maintain tissue homeostasis. **a**, In the healthy HDAC3-sufficient intestine, HDAC3-dependent maintenance of intestinal homeostasis reflects an integrated effect of commensal-derived signals and host transcriptional networks. **b**, Impaired

IEC-intrinsic HDAC3-dependent gene regulation results in increased IEC proliferation, altered Paneth cell survival, intestinal dysbiosis, impaired intestinal barrier function and increased susceptibility to intestinal damage and inflammation.



Extended Data Figure 10 | Pyrosequencing parameters. **a**, Number of reads and alpha diversity (observed species). **b**, Rarefaction curves. **c**, Principal coordinate analysis 2D plots. **d**, Hierarchical clustering dendrograms.

Immunosuppressive CD71⁺ erythroid cells compromise neonatal host defence against infection

Shokrollah Elahi¹, James M. Ertel¹, Jeremy M. Kinder¹, Tony T. Jiang¹, Xuzhe Zhang¹, Lijun Xin¹, Vandana Chaturvedi¹, Beverly S. Strong², Joseph E. Qualls¹, Kris A. Steinbrecher³, Theodosia A. Kalfa⁴, Aimen F. Shaaban² & Sing Sing Way¹

Newborn infants are highly susceptible to infection. This defect in host defence has generally been ascribed to the immaturity of neonatal immune cells; however, the degree of hyporesponsiveness is highly variable and depends on the stimulation conditions^{1–7}. These discordant responses illustrate the need for a more unified explanation for why immunity is compromised in neonates. Here we show that physiologically enriched CD71⁺ erythroid cells in neonatal mice and human cord blood have distinctive immunosuppressive properties. The production of innate immune protective cytokines by adult cells is diminished after transfer to neonatal mice or after co-culture with neonatal splenocytes. Neonatal CD71⁺ cells express the enzyme arginase-2, and arginase activity is essential for the immunosuppressive properties of these cells because molecular inhibition of this enzyme or supplementation with L-arginine overrides immunosuppression. In addition, the ablation of CD71⁺ cells in neonatal mice, or the decline in number of these cells as postnatal development progresses parallels the loss of suppression, and restored resistance to the perinatal pathogens *Listeria monocytogenes* and *Escherichia coli*^{8,9}. However, CD71⁺ cell-mediated susceptibility to infection is counterbalanced by CD71⁺ cell-mediated protection against aberrant immune cell activation in the intestine, where colonization with commensal microorganisms occurs swiftly after parturition^{10,11}. Conversely, circumventing such colonization by using antimicrobials or gnotobiotic germ-free mice overrides these protective benefits. Thus, CD71⁺ cells quench the excessive inflammation induced by abrupt colonization with commensal microorganisms after parturition. This finding challenges the idea that the susceptibility of neonates to infection reflects immune-cell-intrinsic defects and instead highlights processes that are developmentally more essential and inadvertently mitigate innate immune protection. We anticipate that these results will spark renewed investigation into the need for immunosuppression in neonates, as well as improved strategies for augmenting host defence in this vulnerable population.

Neonates are highly susceptible to disseminated infections, which are often fatal. Numerous distinctions have been described between neonatal and adult responses to infection, including blunted inflammatory cytokine production, skewed T-helper-cell differentiation and fewer protective immune cells; however, the degree of neonatal immune cell hyporesponsiveness varies markedly with the stimulation conditions^{1–7}. Thus, given that neonatal cells have the potential for activation, a more unified explanation is needed for why neonates remain susceptible to infection. We found that the susceptibility of human neonates to infection with the bacterium *L. monocytogenes* is recapitulated in neonatal mice^{8,12} (Fig. 1a). Given the delayed immunological development in mice at birth^{7,13}, 6-day-old mice were used as neonates, and their responses were compared with 8-week-old (adult) mice. In addition to diminished survival, over 1,000-fold more *L. monocytogenes* bacteria were recovered from neonatal mice than from adult mice, and this lack of susceptibility in adults was maintained after adjusting the bacterial inoculation dose

proportionally to increased body weight (Fig. 1b). Accordingly, neonatal mice, like newborn humans, are intrinsically susceptible to disseminated infection.

To investigate the cellular basis of neonatal susceptibility, the effect of adoptively transferring immune cells from adult mice was evaluated (Fig. 1c and Extended Data Fig. 1a). We reasoned that if neonatal susceptibility reflects an inadequate number or a hyporesponsiveness of immune cells, then transferred adult cells would restore protection. However, neonates containing adult splenocytes remained equally susceptible to *L. monocytogenes* infection (Fig. 1d). Given these somewhat surprising results, the activation of adult cells within neonates was investigated. Because differences in susceptibility between neonatal and adult mice become apparent within 48 h of infection (Fig. 1b), we focused on essential innate immune protective cytokines such as tumour-necrosis factor- α (TNF- α)^{14–16}. Remarkably, when adult splenocytes containing CD11b⁺ granulocyte/macrophage cells, CD11c⁺ dendritic cells or B220⁺ lymphocytes were transferred to neonatal mice, their TNF- α production induced by *L. monocytogenes* infection was extinguished to levels comparable to that of endogenous neonatal cells (Fig. 1e and Extended Data Fig. 1b). Conversely, TNF- α production by neonatal cells was restored after transfer to *L. monocytogenes*-infected adult mice (Extended Data Fig. 1c). These findings suggest that neonatal infection susceptibility might not simply reflect immune-cell-intrinsic defects but instead active suppression within the neonatal environment.

To assess the potential immunosuppressive properties of neonatal cells, the activation and cytokine production of adult immune cells co-cultured with neonatal splenocytes were evaluated. Consistent with the diminished responsiveness of neonatal cells to purified microbial ligands^{1,3,5,6}, these cells produced considerably less TNF- α and interleukin-6 (IL-6) after stimulation with heat-killed *L. monocytogenes* than did adult mouse splenocytes (Fig. 1f). Similar defects were found for human cord blood cells compared with adult peripheral blood mononuclear cells (Extended Data Fig. 2). Interestingly, combining neonatal and adult splenocytes caused a precipitous decline in cytokine production compared with cultures containing only adult cells (Fig. 1f). Varying the number of neonatal splenocytes in the presence of a fixed quantity of adult cells identified by expression of the congenic marker CD45.1 showed that TNF- α production by adult CD11b⁺, CD11c⁺ or B220⁺ cells was restricted in a dose-dependent manner (Fig. 2a and Extended Data Fig. 3a). Immunosuppression also extended to T cells because neonatal splenocytes impeded the upregulation of early activation markers (such as CD69 and CD25) among adult CD8⁺ cells after anti-CD3 antibody stimulation (Fig. 2a and Extended Data Fig. 3b). Thus, neonatal splenocytes have suppressive properties that recapitulate the blunted activation of adult immune cells within infected neonates.

To establish the molecular basis by which neonatal cells mediate suppression, the effect of inhibitors or neutralizing antibodies on immunomodulatory pathways was evaluated in co-culture. We found that overriding the enzymatic activity of arginase by addition of the inhibitors

¹Division of Infectious Diseases and Perinatal Institute, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA. ²Center for Fetal Cellular and Molecular Therapy, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA. ³Division of Gastroenterology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA. ⁴Division of Hematology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA.

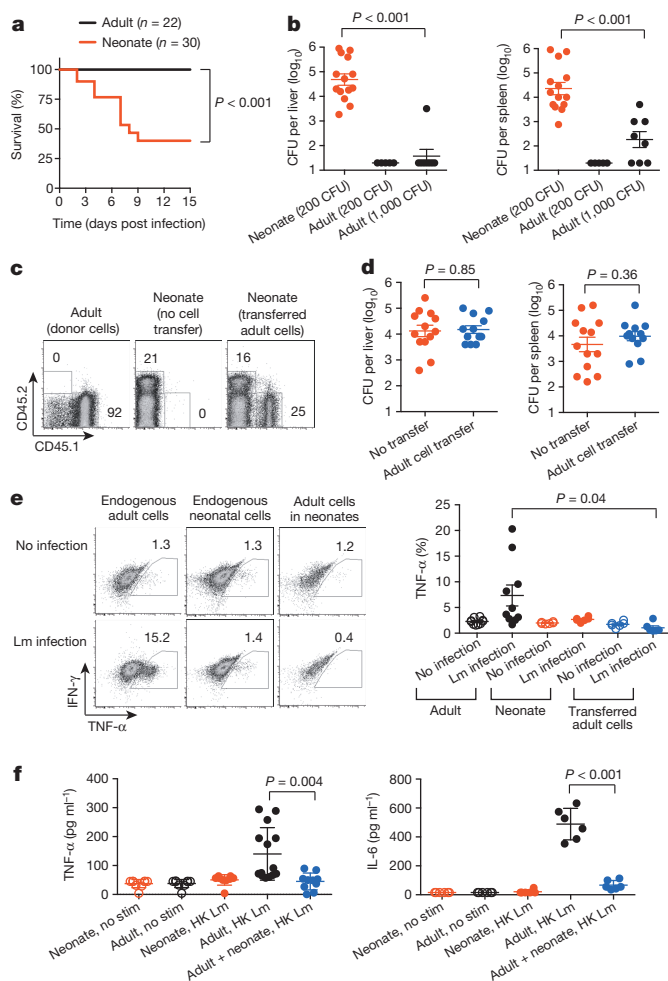


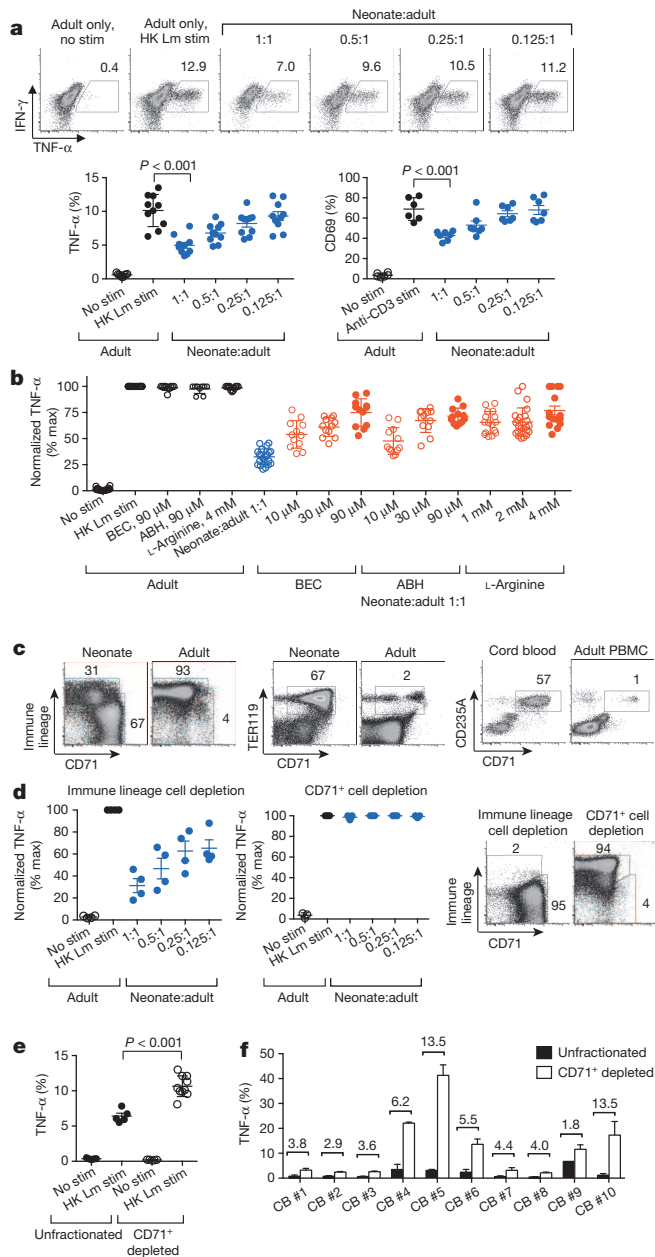
Figure 1 | Infection susceptibility of neonatal mice and immunosuppressive properties of neonatal cells. **a**, Survival of 6-day-old (neonatal) and 8-week-old (adult) mice after *L. monocytogenes* (Lm) infection (200 colony-forming units (CFU)). **b**, Number of recoverable bacteria 48 h after infection with various doses of *L. monocytogenes*. **c**, Flow cytometric analysis showing the retained donor CD45.1⁺ cells from adult mice in the splenocyte population of neonatal mice. Numbers indicate the percentage of cells in the adjacent boxed areas. **d**, Number of recoverable bacteria after *L. monocytogenes* infection (200 CFU) of neonatal mice containing donor adult splenocytes. **e**, Representative flow cytometry plots (left) and cumulative composite data (right) for the percentage of endogenous CD11b⁺ adult and neonatal cells, and CD11b⁺ donor adult cells within neonates, that produce TNF- α *ex vivo* 48 hours after *L. monocytogenes* infection. **f**, Cytokine production by neonatal or adult mouse splenocytes after stimulation (stim) with heat-killed (HK) *L. monocytogenes* individually or in co-culture for 72 h as measured by enzyme-linked immunosorbent assay (ELISA). Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m. IFN- γ , interferon- γ .

BEC, ABH, nor-NOHA or L-NOHA, or by supplementation with L-arginine, restored the activation of adult responder cells co-cultured with neonatal splenocytes (Fig. 2b and Extended Data Fig. 4). By contrast, the inhibition of other immunomodulatory molecules, including indoleamine 2,3-dioxygenase (IDO), transforming growth factor- β (TGF- β), IL-10 or reactive oxygen species, had no significant effect (Extended Data Fig. 4). Importantly, arginase inhibition did not influence TNF- α production by cultures containing only adult cells, illustrating that restored cytokine production was the result of reversing the suppression by neonatal cells. Thus, neonatal splenocytes control immune cell activation through arginine depletion, similarly to the suppressor cells that are associated with tumour progression or persistent infection^{17,18}.

We next sought to define the neonatal splenocyte subset that is responsible for suppression. In contrast to adult mouse splenocytes, $\sim 95\%$ of which express immune lineage markers (CD4, CD8, CD11b, CD11c, B220 and NK1.1), this combination of markers was found on $< 35\%$ of neonatal splenocytes (Fig. 2c). Nearly all of the remaining neonatal splenocytes co-expressed the transferrin receptor (CD71) and the erythroid-lineage-defining molecule TER119 (also known as Ly-76)^{19,20}. Similarly, human cord blood cells contain an equally enriched proportion of CD71⁺ cells that co-express the human erythroid marker CD235A, which is in contrast to adult peripheral blood mononuclear cells (Fig. 2c and Extended Data Fig. 5a). To establish which subset confers suppression, neonatal splenocytes were fractionated using anti-CD71 antibody or a cocktail of anti-immune-lineage antibodies, and the suppressive properties of each cell population were evaluated in co-culture with adult responder splenocytes. The depletion of CD71⁺ cells eliminated the suppression. By contrast, the depletion of immune lineage cells not only retained but also exaggerated the suppression by the remaining CD71⁺ cells (Fig. 2d). Moreover, the depletion of CD71⁺ cells in mouse splenocyte or human cord blood cell populations unleashed more robust cytokine production by the remaining immune lineage cells, indicating that CD71⁺ cells also impair neonatal immune cell activation (Fig. 2e, f and Extended Data Figs 5b and 6). Likewise, *L. monocytogenes* infection resulted in greater activation of the immune cells recovered from neonatal mouse lymph nodes, where CD71⁺ cells are naturally present only in small numbers (Extended Data Fig. 7). Thus, enriched CD71⁺ erythroid cells in neonates suppress systemic immune cell activation.

CD71⁺ cell ablation was used to further establish the relationship between suppression by these cells and neonatal susceptibility to infection. Although only $\sim 60\%$ depletion could be achieved, significant reductions in the number of recoverable *L. monocytogenes* bacteria were found after anti-CD71 antibody treatment compared with isotype antibody treatment (Fig. 3a). The protective benefits of CD71⁺ cell depletion in neonates similarly extended to *E. coli* infection (Extended Data Fig. 8). Likewise, the physiological contraction of the CD71⁺ cell population as postnatal development progressed paralleled the gradual restoration of protection against infection to adult levels. At 6 and 9 days after parturition, mice had equally high CD71⁺ cell numbers and comparable numbers of recoverable bacteria after infection. By contrast, 15-day-old mice had $\sim 60\%$ fewer CD71⁺ cells and a 100-fold lower pathogen burden. In 21-day-old mice, CD71⁺ cell numbers had declined to levels comparable to those of adult mice, and pathogen burden was undetectable (Fig. 3b, c). The progressive reduction in susceptibility to infection with the decline in CD71⁺ cell numbers (as postnatal development progressed) also paralleled the loss of immunosuppressive properties among unfractionated splenocytes (Fig. 3d). Thus, enriched CD71⁺ cells dictate neonatal infection susceptibility, because protection is restored by antibody-mediated depletion of these cells or by their natural disappearance during postnatal development.

To determine whether suppression by CD71⁺ cells extends beyond the neonatal period, infection susceptibility after phlebotomy-induced anaemia, with ensuing erythroid cell population expansion, was evaluated in adult mice. Although anaemia efficiently induced CD71⁺TER119⁺ cell accumulation, with subsets comparable to those of neonatal splenocytes²¹, there were no significant shifts in infection susceptibility (Extended Data Fig. 9), and purified adult CD71⁺ cells showed no immunosuppressive properties in co-culture assays (Fig. 3e). The lack of suppression by adult CD71⁺ cells paralleled the markedly lower arginase-2 expression in these cells than in neonatal CD71⁺ cells (Fig. 3f) and the necessity for arginase enzymatic activity for suppression by neonatal splenocytes (Fig. 2b). Thus, suppression by CD71⁺ cells is likely to be restricted to the population that is naturally enriched in neonates. These findings are consistent with the lack of susceptibility to infection among individuals with ailments that accelerate erythropoiesis (for example, thalassaemia and spherocytosis) after the neonatal period, as well as the immunomodulatory properties of other



erythroid cell subsets and the remarkable diversity in gene expression among neonatal erythroid precursor cells compared with adult erythroid precursor cells^{22–28}.

Although these results establish that CD71⁺ cells impair neonatal host defence against infection, they also raise exciting new questions about why suppressive cells are temporally enriched in neonates. One possibility is to avert the excessive inflammation that might otherwise occur with the abrupt transition from a sterile *in utero* setting to colonization with commensal microbes in the external environment²⁹. This idea is supported by the finding that intestinal immune cells in neonatal mice are selectively activated by anti-CD71 antibody but not by isotype control antibody, as well as by the rapid, intensifying colonization of the intestine with commensal microbes in the first week after birth^{10,11} (Fig. 4a–c). In particular, intestinal CD11b⁺ and CD11c⁺ cells from CD71⁺ cell-depleted neonatal mice produce significantly more TNF-α and upregulate expression of the co-stimulatory molecules CD40, CD80 and CD86 more than analogous cells from neonatal mice treated with isotype control antibody (Fig. 4a, b and Extended Data Fig. 10a). By contrast, after CD71⁺ cell depletion, these activation parameters did not change in cell populations from the spleen or

Figure 2 | Arginase inhibition overrides immunosuppression by neonatal splenocytes containing enriched CD71⁺ erythroid cells. **a**, Representative flow cytometry plots (top) and cumulative composite data (bottom) showing TNF-α production by adult CD11b⁺ cells co-cultured with neonatal splenocytes in the indicated ratios and stimulated (stim) with heat-killed (HK) *L. monocytogenes* (Lm) (bottom left), as well as CD69 expression among adult CD8⁺ cells co-cultured with neonatal splenocytes in the indicated ratios and stimulated with anti-CD3 antibody (bottom right). **b**, Normalized (percentage maximum, % max) TNF-α production by CD11b⁺ adult cells cultured alone (black) or co-cultured with neonatal splenocytes at a 1:1 ratio (blue), or additional supplementation (red) with the arginase inhibitors BEC or ABH, or L-arginine. **c**, Flow cytometric analysis showing the proportion of immune lineage (CD4⁺CD8⁺CD11b⁺CD11c⁺B220⁺NK1.1⁺) cells or CD71⁺TER119⁺ erythroid lineage cells in neonatal mouse splenocyte populations compared with adult mouse splenocyte populations, or the proportion of CD71⁺CD235A⁺ cells in human cord blood cell populations compared with adult peripheral blood mononuclear cell (PBMC) populations. **d**, TNF-α production by adult CD11b⁺ cells co-cultured with immune-lineage-cell-depleted or CD71⁺ cell-depleted neonatal splenocytes (left), and representative flow cytometry plots illustrating the efficiency of cell depletion (right). **e**, TNF-α production by unfractionated or CD71⁺ cell-depleted neonatal CD11b⁺ cells. **f**, TNF-α production by unfractionated or CD71⁺ cell-depleted CD11b⁺ human cord blood cells. The fold increase in TNF-α production by the CD71⁺ cell-depleted cell populations compared with the unfractionated controls is shown within the graph. These results are representative of three independent experiments for mice and ten individual cord blood (CB) samples. Error bars, mean ± s.e.m.

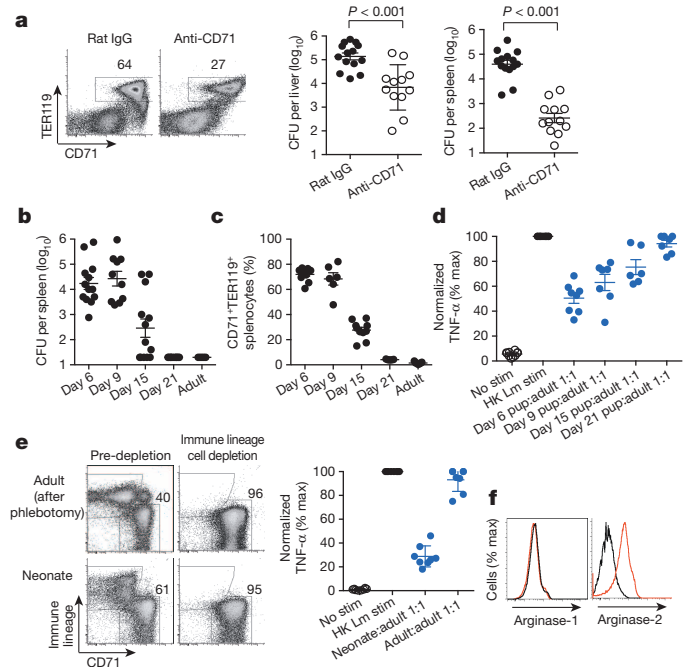


Figure 3 | Enriched CD71⁺ cells compromise neonatal host defence against infection. **a**, The proportion of CD71⁺TER119⁺ splenocytes and the number of recoverable bacteria (CFU) 48 h after *L. monocytogenes* (Lm) infection of neonatal mice treated with anti-CD71 antibody or isotype control (rat IgG) antibody. **b**, The number of recoverable bacteria after *L. monocytogenes* infection of mice during postnatal development and adult mice. **c**, The proportion of CD71⁺TER119⁺ splenocytes in mice during postnatal development and adult mice. **d**, Normalized TNF-α production (percentage maximum, % max) by adult CD11b⁺ responder cells co-cultured with a 1:1 ratio of splenocytes from mice in each age group. **e**, Representative flow cytometry plots showing the efficiency of immune lineage cell depletion (left), and cumulative composite data for TNF-α production by adult CD11b⁺ cells co-cultured with purified CD71⁺ cells from neonatal or phlebotomized adult mice (right). **f**, Arginase-1 and arginase-2 expression in CD71⁺ cells from phlebotomized adult (black) or neonatal (red) splenocytes. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean ± s.e.m. HK, heat-killed; stim, stimulation.

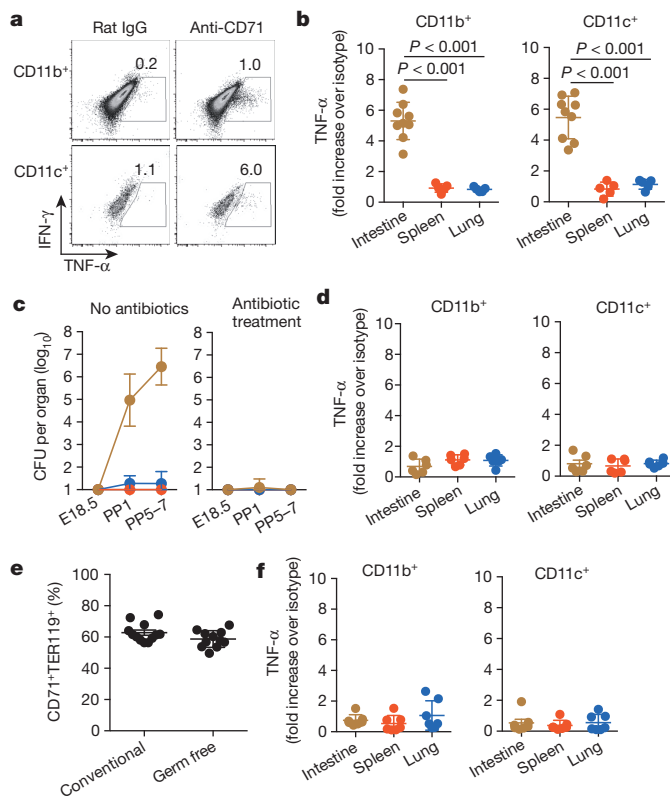


Figure 4 | Neonatal CD71⁺ cells prevent aberrant immune cell activation in tissue that is rapidly colonized with commensal microbes. **a**, TNF- α production by intestinal immune cells from 8-day-old mice treated with anti-CD71 antibody (or isotype control antibody (rat IgG)) on days 5 and 6. **b**, TNF- α production by cells recovered from the indicated tissues after anti-CD71 antibody treatment of neonatal mice (normalized to the values from isotype-control-antibody-treated neonatal mice). **c**, The number of recoverable bacteria (CFU) from intestine (brown), spleen (red) and lung (blue) of fetal (embryonic day (E) 18.5) and neonatal mice in the indicated age (days post parturition (PP)) after housing with unsupplemented drinking water or drinking water containing antibiotics (ampicillin, gentamicin, metronidazole, neomycin and vancomycin) from E14.5 ($n = 6$ –18 mice per time point). **d**, TNF- α production by cells recovered from the indicated tissues of neonatal mice sustained on antimicrobial therapy and treated with anti-CD71 antibody (normalized to the values from isotype-control-antibody-treated neonatal mice sustained on antimicrobial therapy). **e**, The proportion of splenocytes that are CD71⁺TER119⁺ in 8-day-old conventional and germ-free mice. **f**, TNF- α production by cells recovered from the indicated tissues of germ-free neonatal mice treated with anti-CD71 antibody (normalized to the values from isotype-control-antibody-treated germ-free neonatal mice). Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.

lung, which remain sterile or become colonized with considerably fewer commensal bacteria than the intestine (Fig. 4b, c and Extended Data Fig. 10a).

Antimicrobials were used to further investigate the relationship between commensal bacteria and CD71⁺ cell-mediated protection against aberrant intestinal immune cell activation. We found that a defined antibiotic cocktail that eradicated intestinal bacteria when administered to the drinking water of pregnant mice also prevented colonization in neonates³⁰ (Fig. 4c). Eliminating commensal bacteria, in turn, abolished the increase in TNF- α production and co-stimulatory molecule expression induced by CD71⁺ cell depletion (Fig. 4d and Extended Data Fig. 10b). To more definitively establish the necessity for commensal microbes in promoting intestinal inflammation, cell activation induced by CD71⁺ cell depletion was further addressed using germ-free mice. Although CD71⁺ cells were equally enriched in gnotobiotic neonatal mice and conventional neonatal mice, their depletion from gnotobiotic

germ-free mice did not induce significant changes in intestinal immune cell activation, similarly to antibiotic treatment (Fig. 4e, f and Extended Data Fig. 10c). Thus, neonatal CD71⁺ erythroid cells protect against excessive inflammation triggered by commensal microbes, whereas their elimination (by using antimicrobials or under germ-free conditions) alleviates these protective benefits.

Taken together, our results demonstrate that neonatal infection susceptibility results from the temporal presence of enriched immunosuppressive CD71⁺ cells and is an unfortunate by-product of the greater benefits of active suppression during this crucial developmental period, when tolerance to commensal microbes is more uniformly advantageous. We anticipate that these findings will spur renewed investigation of why neonatal protection against infection is compromised, as well as the study of therapeutic approaches aimed at dissociating the beneficial and harmful effects of CD71⁺ cells for augmenting host defence in this vulnerable population.

METHODS SUMMARY

Pregnant C57BL/6 mice were checked twice daily for birth timing. For infection, mice were inoculated intraperitoneally with *L. monocytogenes* or *E. coli*, and susceptibility was determined by counting the number of CFU arising from dilutions of organ homogenate spread onto agar plates. For cell transfer, splenocytes from congenic donors were injected into recipients 1 day before infection. For co-culture, adult splenocytes were seeded with neonatal cells and stimulated with either heat-killed *L. monocytogenes* or anti-CD3 antibody. For depletion, 150 μ g anti-CD71 antibody (8D3, AbD Serotec) or isotype control antibody was administered on days 5 and 6 after birth. Differences in survival were compared using the Mantel–Cox log-rank test. Differences in the number of CFU, the cytokine levels and the cell activation levels between separate groups or related groups were analysed using an unpaired or paired Student's *t*-test, respectively.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 March; accepted 18 September 2013.

Published online 6 November 2013.

- Kollmann, T. R., Levy, O., Montgomery, R. R. & Goriely, S. Innate immune function by Toll-like receptors: distinct responses in newborns and the elderly. *Immunity* **37**, 771–783 (2012).
- PrabhuDas, M. *et al.* Challenges in infant immunity: implications for responses to infection and vaccines. *Nature Immunol.* **12**, 189–194 (2011).
- Zaghuan, H., Hoeman, C. M. & Adkins, B. Neonatal immunity: faulty T-helpers and the shortcomings of dendritic cells. *Trends Immunol.* **30**, 585–591 (2009).
- Adkins, B., Leclerc, C. & Marshall-Clarke, S. Neonatal adaptive immunity comes of age. *Nature Rev. Immunol.* **4**, 553–564 (2004).
- Kollmann, T. R. *et al.* Neonatal innate TLR-mediated responses are distinct from those of adults. *J. Immunol.* **183**, 7150–7160 (2009).
- Levy, O. *et al.* Selective impairment of TLR-mediated innate immunity in human newborns: neonatal blood plasma reduces monocyte TNF- α induction by bacterial lipopeptides, lipopolysaccharide, and imiquimod, but preserves the response to R-848. *J. Immunol.* **173**, 4627–4634 (2004).
- Siegrist, C. A. Neonatal and early life vaccinology. *Vaccine* **19**, 3331–3346 (2001).
- Gellin, B. G. & Broome, C. V. Listeriosis. *J. Am. Med. Assoc.* **261**, 1313–1320 (1989).
- Camacho-Gonzalez, A., Spearman, P. W. & Stoll, B. J. Neonatal infectious diseases: evaluation of neonatal sepsis. *Pediatr. Clin. North Am.* **60**, 367–389 (2013).
- Rotimi, V. O. & Duerden, B. I. The development of the bacterial flora in normal neonates. *J. Med. Microbiol.* **14**, 51–62 (1981).
- Ducluzeau, R. Implantation and development of the gut flora in the newborn animal. *Ann. Rech. Vet.* **14**, 354–359 (1983).
- Wirsing von König, C. H., Heymer, B., Finger, H., Emmerling, P. & Hof, H. Alteration of non-specific resistance to infection with *Listeria monocytogenes*. *Infection* **16** (suppl. 2), S112–S117 (1988).
- Mold, J. E. & McCune, J. M. Immunological tolerance during fetal development: from mouse to man. *Adv. Immunol.* **115**, 73–111 (2012).
- Havell, E. A. Evidence that tumor necrosis factor has an important role in antibacterial resistance. *J. Immunol.* **143**, 2894–2899 (1989).
- Nakane, A., Minagawa, T. & Kato, K. Endogenous tumor necrosis factor (cachectin) is essential to host resistance against *Listeria monocytogenes* infection. *Infect. Immun.* **56**, 2563–2569 (1988).
- Pasparakis, M., Alexopoulou, L., Episkopou, V. & Kollias, G. Immune and inflammatory responses in TNF- α -deficient mice: a critical requirement for TNF- α in the formation of primary B cell follicles, follicular dendritic cell networks and germinal centers, and in the maturation of the humoral immune response. *J. Exp. Med.* **184**, 1397–1411 (1996).
- Bronte, V. & Zanovello, P. Regulation of immune responses by L-arginine metabolism. *Nature Rev. Immunol.* **5**, 641–654 (2005).

18. Morris, S. M. Jr. Arginine: master and commander in innate immune responses. *Sci. Signal.* **3**, pe27 (2010).
19. Hermansen, M. C. Nucleated red blood cells in the fetus and newborn. *Arch. Dis. Child. Fetal Neonatal Ed.* **84**, F211–F215 (2001).
20. Opiela, S. J., Levy, R. B. & Adkins, B. Murine neonates develop vigorous *in vivo* cytotoxic and T_H1/T_H2 responses upon exposure to low doses of NIMA-like alloantigens. *Blood* **112**, 1530–1538 (2008).
21. Kalfa, T. A. *et al.* Rac1 and Rac2 GTPases are necessary for early erythropoietic expansion in the bone marrow but not in the spleen. *Haematologica* **95**, 27–35 (2010).
22. Prins, H. A. *et al.* Arginase release from red blood cells: possible link in transfusion induced immune suppression? *Shock* **16**, 113–115 (2001).
23. Millington, O. R., Di Lorenzo, C., Phillips, R. S., Garside, P. & Brewer, J. M. Suppression of adaptive immunity to heterologous antigens during *Plasmodium* infection through hemozoin-induced failure of dendritic cell function. *J. Biol.* **5**, 5 (2006).
24. Muszynski, J. *et al.* Immunosuppressive effects of red blood cells on monocytes are related to both storage time and storage solution. *Transfusion* **52**, 794–802 (2012).
25. Morera, D. & Mackenzie, S. A. Is there a direct role for erythrocytes in the immune response? *Vet. Res.* **42**, 89 (2011).
26. Jackson, A., Nanton, M. R., O'Donnell, H., Akue, A. D. & McSorley, S. J. Innate immune activation during *Salmonella* infection initiates extramedullary erythropoiesis and splenomegaly. *J. Immunol.* **185**, 6198–6204 (2010).
27. Akinosoglou, K. S., Solomou, E. E. & Gogos, C. A. Malaria: a haematological disease. *Hematology* **17**, 106–114 (2012).
28. Kingsley, P. D. *et al.* Ontogeny of erythroid gene expression. *Blood* **121**, e5–e13 (2013).
29. Maynard, C. L., Elson, C. O., Hatton, R. D. & Weaver, C. T. Reciprocal interactions of the intestinal microbiota and immune system. *Nature* **489**, 231–241 (2012).
30. Abt, M. C. *et al.* Commensal bacteria calibrate the activation threshold of innate antiviral immunity. *Immunity* **37**, 158–170 (2012).

Acknowledgements We thank D. Haslam, M. Hostetter, L. Muglia, J. Whitsett and C. Wilson for discussions, J. Mortensen for help with anaerobic cultures, and K. Eaton, C. Schray and the University of Michigan Unit for Laboratory Animal Medicine for providing germ-free mice. We thank the Mount Auburn OB-GYN associates, OB-GYN residents, and the University and Christ Hospital labour and delivery nursing staff for collecting cord blood, the Cell Processing and Manipulation Core for obtaining peripheral blood from adult donors, and the CCHMC Translational Research Trials Office for providing the regulatory and administrative support for studies with human cells. This research was supported by NIAID (R01AI087830 and R01AI100934) (S.S.W.) and NHLBI (R01HL103745) (A.F.S.). S.S.W. holds an Investigator in the Pathogenesis of Infectious Disease award from the Burroughs Wellcome Fund.

Author Contributions All authors performed or participated in the design of the experiments. S.E. and S.S.W. wrote the paper with editorial input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S.W. (singsing.way@cchmc.org).

METHODS

Mice. C57BL/6 ($CD45.2^{+}CD45.1^{-}$) and congenic $CD45.1^{+}CD45.2^{-}$ mice were purchased from the National Cancer Institute and checked twice daily for pregnancy and birth timing. For infection and immune cell analysis, male and female mice were used in all age groups tested. For phlebotomy, mice were bled (300–350 μ l, with intraperitoneal saline replacement) for 3 consecutive days and analysed for 3 days thereafter. For eradication of intestinal bacteria, autoclaved drinking water was supplemented with 0.5 mg ml⁻¹ ampicillin, 0.5 mg ml⁻¹ gentamicin, 0.5 mg ml⁻¹ metronidazole, 0.5 mg ml⁻¹ neomycin and 0.25 mg ml⁻¹ vancomycin beginning on embryonic day (E) 14.5, and this was maintained throughout parturition and nursing. Gnotobiotic germ-free C57BL/6 mice were maintained by the University of Michigan Unit for Laboratory Animals. All experiments were performed in accordance with Cincinnati Children's Hospital Medical Center and University of Michigan Institutional Animal Care and Use Committee approved protocols.

Infection and enumerating bacterial recovery. *L. monocytogenes* (strain 10403S) or *E. coli* (strain UTI89) was grown in brain heart infusion (BHI) medium at 37 °C, back diluted to early logarithmic phase (an optical density at 600 nm (OD₆₀₀) of 0.1) and resuspended in sterile saline. Mice were inoculated intraperitoneally with a dose of 2×10^2 or 1×10^3 CFU per mouse. The inoculum for each experiment was verified by spreading a diluted aliquot onto agar plates. To assess susceptibility after infection, mouse organs (spleen, liver, lung and intestine) were dissected and homogenized in sterile saline containing 0.05% Triton X-100 to disperse the intracellular bacteria, and serial dilutions of the organ homogenate were spread onto agar plates. Colonies were counted after plate incubation at 37 °C. To evaluate anaerobic growth, organ homogenates were plated onto pre-reduced media (Anaerobe Systems) and incubated for 72 h at 37 °C using a Whitley A35 anaerobic workstation (Don Whitley Scientific).

Antibodies and flow cytometry. Fluorophore- or biotin-conjugated antibodies specific for mouse cell surface antigens and cytokines were purchased from eBioscience or BD Biosciences. The following antibodies were used: anti-B220 (RA3-6B2), anti-CD4 (GK1.5), anti-CD8a (53-6.7), anti-CD11b (M1/70), anti-CD11c (N418), anti-CD25 (PC61.5), anti-CD69 (H1.2F3), anti-CD45.1 (A20), anti-CD45.2 (104), anti-CD71 (R17217 and C2F2), anti-NK1.1 (PK136), anti-CD40 (1C10), anti-CD80 (16-10A1), anti-CD86 (GL1), anti-TER119 (TER-119), anti-IFN- γ (XMGI.2), anti-TNF- α (MP6-XT22), anti-arginase-1 (ARG1-CFS, R&D Systems) and anti-arginase-2 (ab81505, Abcam). For human studies, the following fluorophore- or biotin-conjugated antibodies specific for cell surface markers or cytokines were used: anti-CD3 (UCHT), anti-CD8 (PRA-T8), anti-CD69 (FN50), anti-CD11b (ICRF44), anti-IL-6 (MQ2-13A5), anti-TNF- α (MAB11), anti-CD71 (OKT9) and anti-CD235A. For *in vivo* depletion, 150 μ g purified anti-CD71 (8D3, AbD Serotec) and rat IgG2a isotype control antibody were administered intraperitoneally 5 and 6 days after birth, corresponding to 1 day before infection and the day of infection.

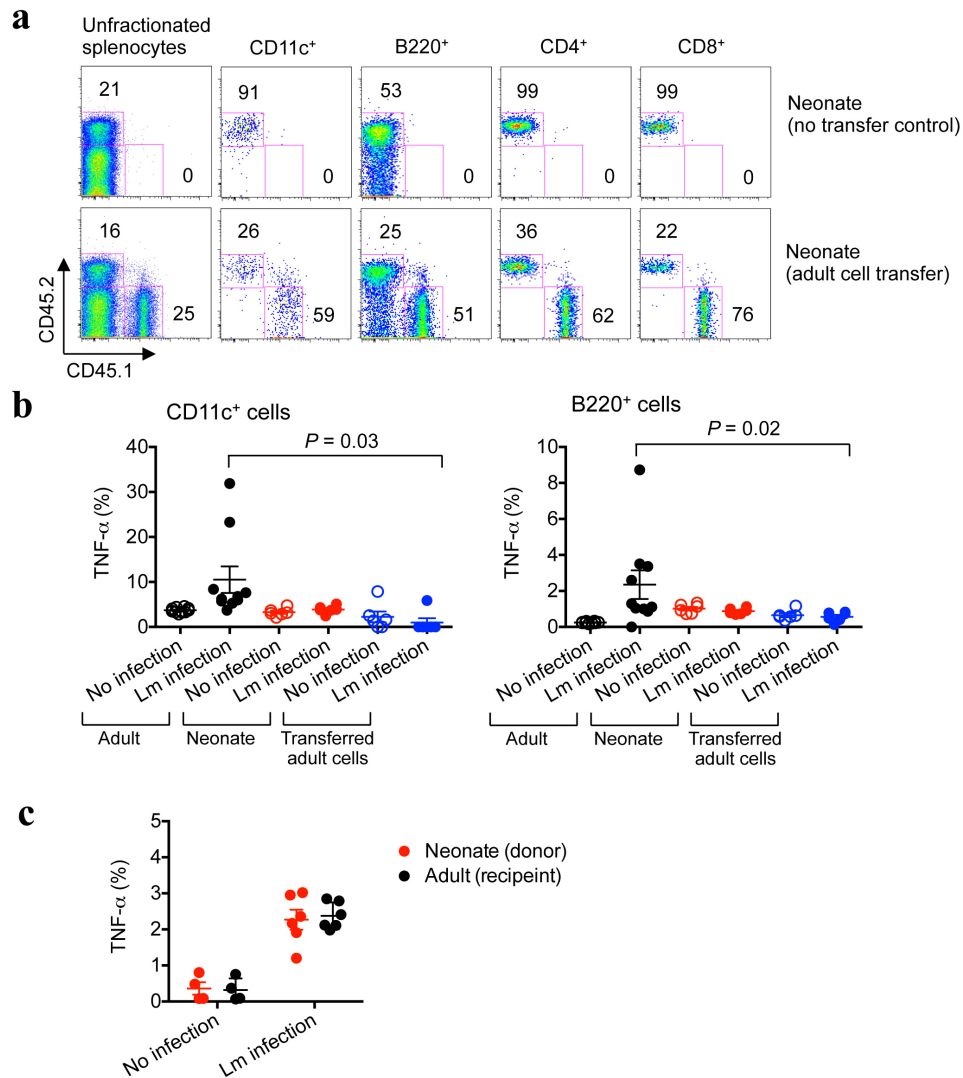
Cell transfer and purification. Splenocytes were collected, and single-cell suspensions were made by grinding between sterile frosted glass slides and filtering through nylon mesh. For adoptive transfer, 5×10^7 splenocytes from congenic donor mice were injected intraperitoneally into recipient mice 1 day before infection. $CD71^{+}$ or immune lineage ($CD4^{+}CD8^{+}CD11b^{+}CD11c^{+}B220^{+}NK1.1^{+}$) cells were purified from neonatal splenocytes by negative selection using biotin-conjugated antibodies and streptavidin-linked magnetic beads (Miltenyi Biotec). For cell isolation from the intestine, the tissue was cut into small pieces in ice-cold Hank's balanced salt solution containing 5 mM EDTA and 1 mM dithiothreitol (DTT) and incubated at 37 °C for 40 min with gentle agitation (220 r.p.m.). The tissue was further minced and digested in medium supplemented with 10% FBS, 500 μ g ml⁻¹ collagenase D, 500 μ g ml⁻¹ DNase and 500 μ g ml⁻¹ dispase and incubated at 37 °C for 60 min with continued gentle agitation. Thereafter, the organ homogenate was strained (100 μ m filter) and separated through a Percoll gradient with centrifugation at 800g for 20 min. For cell isolation from the lung, the tissue was cut into small pieces and incubated in medium supplemented with 10% FBS, 5 mM EDTA, 1 mM DTT, 500 μ g ml⁻¹ DNase I and 500 μ g ml⁻¹ collagenase D for 60 min with gentle agitation (220 r.p.m.). Thereafter, the cell suspension was strained and pelleted by

centrifugation at 450g for 10 min. Immune cells in each sample were identified by staining for the leukocyte common antigens CD45.1 or CD45.2.

Co-culture and stimulation. For *ex vivo* cytokine production, splenocytes were collected 48 h after infection and cultured at 1×10^6 cells ml⁻¹ in DMEM supplemented with 10% FBS and 10 μ g ml⁻¹ brefeldin A for 5 h. For co-culture, a fixed number of responder splenocytes (5×10^5) from $CD45.1^{+}$ adult mice were seeded into 96-well round bottom plates individually or together with $CD45.2^{+}$ neonatal splenocytes at defined ratios and then stimulated for 5 h with 5×10^6 heat-killed *L. monocytogenes* per ml, 0.125 μ g ml⁻¹ anti-mouse CD3 antibody (clone 1C3A1) or 0.125 μ g ml⁻¹ anti-human CD3 antibody (clone UCHT1). Heat-killed *L. monocytogenes* was prepared by growing *L. monocytogenes* 10403S in BHI medium to early logarithmic phase, washing and resuspending the bacteria in sterile saline, and then incubating them at 70 °C for 30 min before storing at -20 °C until use. Each batch was verified as sterile by plating onto BHI agar plates. For intracellular cytokine staining, the medium was supplemented with brefeldin A during stimulation and co-culture. For comparing cytokine production and cell activation between experiments, individual samples were normalized by plotting the percentage maximal response compared with adult cells stimulated without neonatal cells in each experiment. For cytokine production in culture supernatants, 1×10^6 cells ml⁻¹ adult splenocytes or 1×10^6 cells ml⁻¹ neonatal splenocytes were stimulated, separately or together, with 5×10^6 ml⁻¹ heat-killed *L. monocytogenes* for 72 h, and then the concentrations of TNF- α and IL-6 were measured by enzyme-linked immunosorbent assay (ELISA) (R&D Systems). In some experiments, neonatal cells were treated 30 min before and during co-culture with the following: the arginase inhibitors S-(2-boronoethyl)-L-cysteine (BEC), amino-2-borono-6-hexanoic acid (ABH), N-hydroxy-nor-arginine (nor-NOHA) and N-hydroxy-L-arginine (L-NOHA) (each used at 10, 30 and 90 μ M) or 1–4 mM L-arginine; 10–100 μ M apocynin (4'-hydroxy-3'-methoxyacetophenone) to inhibit NADPH oxidase; 100–200 μ M sodium diethyldithiocarbamate trihydrate to inhibit superoxide dismutase; 50–500 μ M 4-aminobenzoic acid hydrazide to inhibit myeloperoxidase; 2–100 μ M SB-431542 hydrate to inhibit the TGF- β receptor; 1–100 μ M Sn(IV) protoporphyrin IX dichloride to inhibit haem oxidation; 100–1,000 μ M 1-methyl-L-tryptophan or 1-methyl-D-tryptophan to inhibit IDO; 1–50 μ g ml⁻¹ anti-IL-10 receptor antibody (clone 1B1); 1–50 μ g ml⁻¹ anti-TGF- β antibody (clone 1D11); or 10–100 μ M N-acetylcysteine.

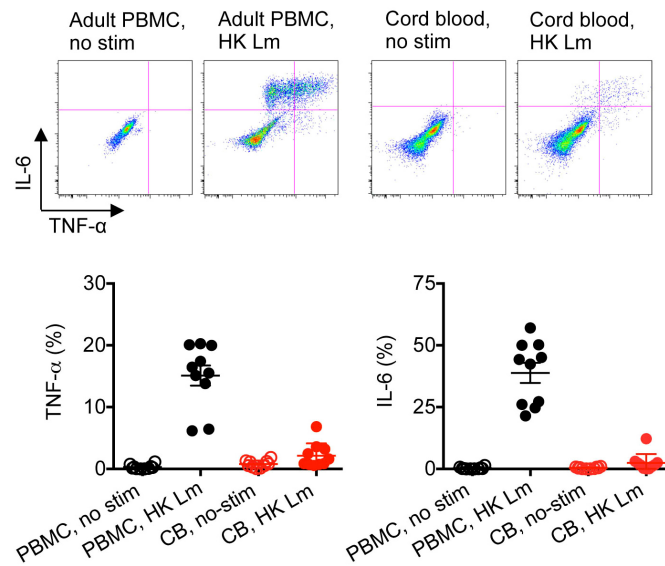
Human sample collection and processing. Peripheral blood was collected from de-identified healthy adult volunteers, and cord blood was collected from term deliveries, under Cincinnati Children's Hospital Medical Center Institutional review board (IRB) approved protocols. Mononuclear cells were freshly isolated over Ficoll-Hypaque gradients. For $CD71^{+}$ cell depletion or mock depletion, cord blood samples were stained using biotin-conjugated anti-CD71 or isotype control antibody and fractionated using streptavidin-linked magnetic beads. For enumerating activation, peripheral or cord blood mononuclear cells were seeded into 96-well round bottom plates (1×10^6 cells per well) in medium supplemented with 10% FBS and stimulated for 5 h with 5×10^6 ml⁻¹ heat-killed *L. monocytogenes* or 0.125 μ g ml⁻¹ anti-CD3 antibody.

Statistical analysis. A sample size of ten per group was planned for each experiment, providing 90% power to detect a 50% difference in response. Depending on the initial results, the sample size in subsequent experiments was adjusted accordingly. Within all figures, each data point reflects results from a single mouse or individual human sample plated in triplicate wells. Neonatal mice in litters were randomized to receive either anti-CD71 or isotype control antibody. For at least one replicate experiment, the investigator was blinded to the treatment of individual groups of neonates. Differences in survival between adult and neonatal mice after infection were compared using the Mantel-Cox log-rank test. The distribution of log₁₀ CFU and the cytokine and cell activation levels were first determined to be normally distributed. Thereafter, differences between separate groups or related groups were analysed using an unpaired or paired Student's *t*-test, respectively, with *P* < 0.05 taken as statistical significance.

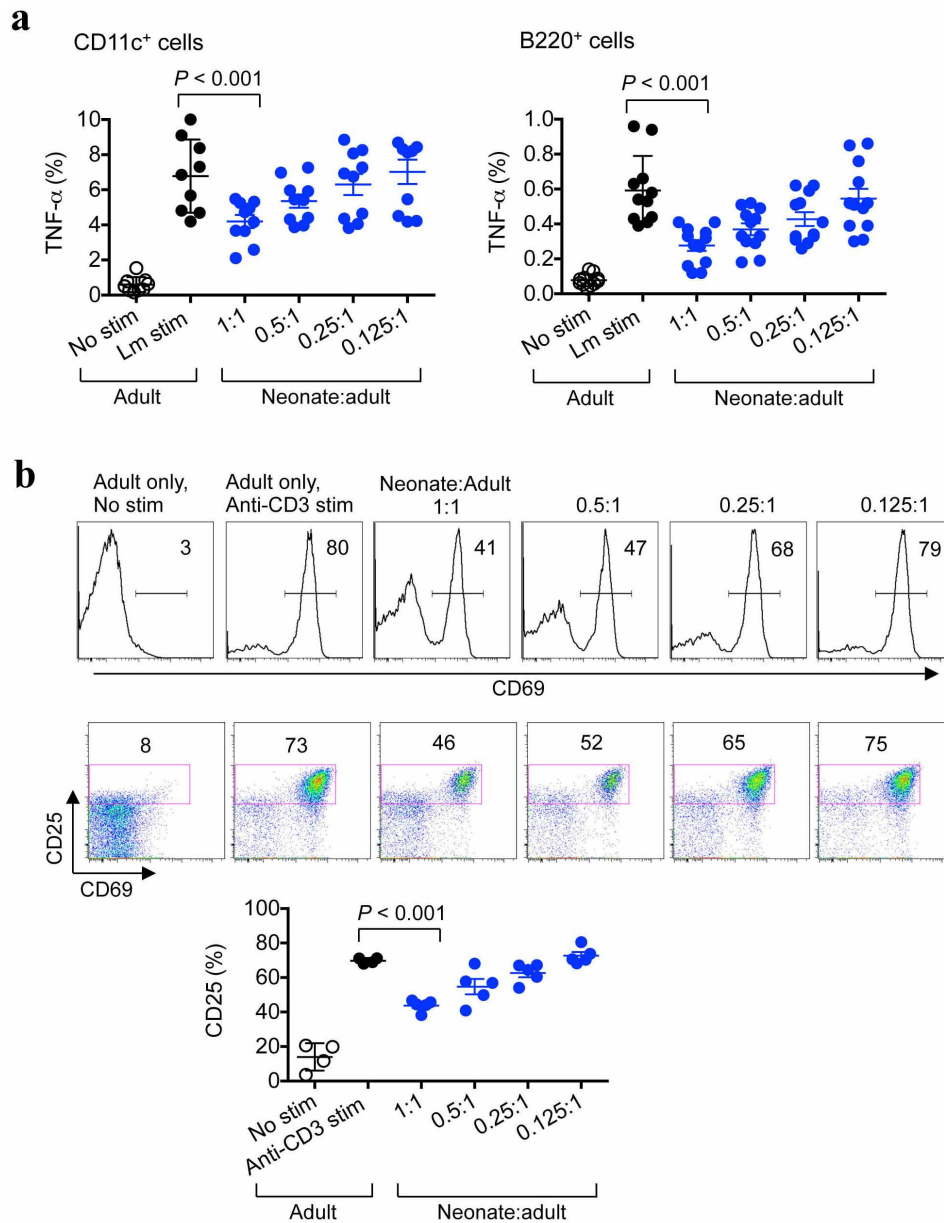


Extended Data Figure 1 | Adoptively transferred splenocyte cells from adult mice are retained but do not become activated in *L. monocytogenes* (Lm)-infected neonatal mice. **a**, The proportion of adult CD45.1⁺ donor cells compared with endogenous neonatal CD45.2⁺ cells among unfractionated splenocytes or various cell subsets 48 h after transfer to 5-day-old neonatal mice. **b**, The proportion of TNF- α -producing CD11c⁺ or B220⁺ cells among endogenous cells in adult or neonatal mice, or among transferred adult cells within neonates. **c**, Restored TNF- α production among neonatal cells after

transfer to adult mice. The proportion of TNF- α -producing CD11b⁺ cells among transferred neonatal cells compared with endogenous adult cells. Forty-eight hours after infection, splenocytes were collected from infected or control mice, cultured in medium containing brefeldin A for 5 h and then subjected to cell surface and intracellular cytokine staining. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.

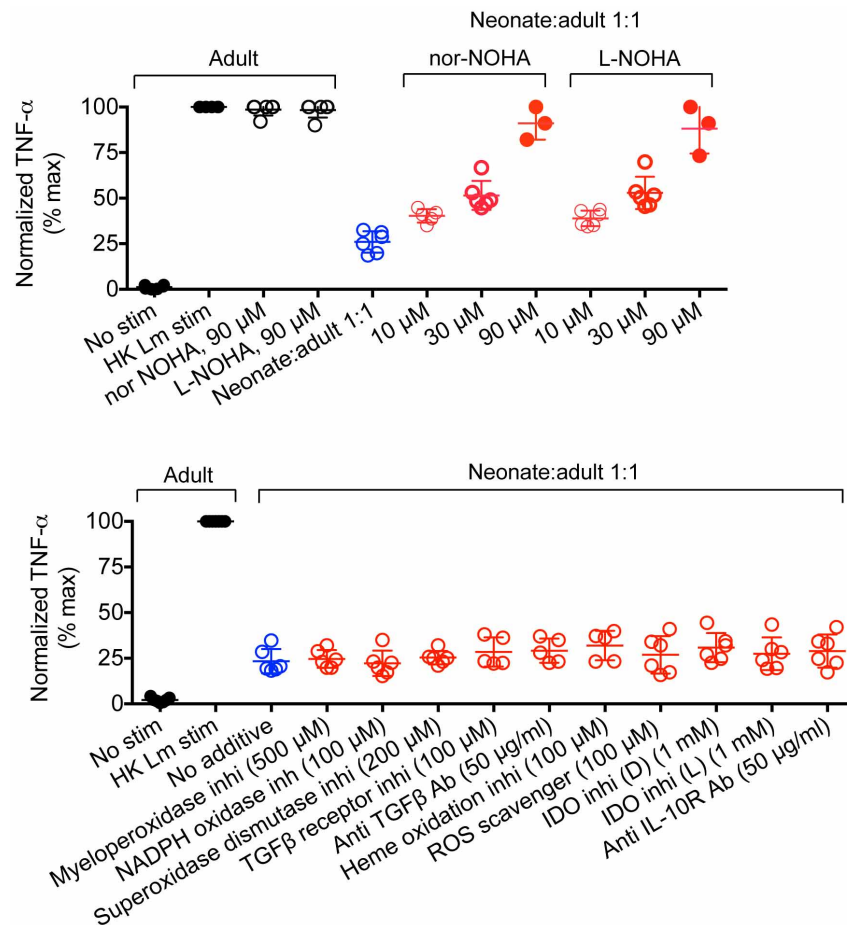


Extended Data Figure 2 | Diminished TNF- α and IL-6 production among human cord blood cells compared with adult peripheral blood mononuclear cells. Representative plots and cumulative data showing TNF- α and IL-6 production by CD11b⁺ cells among adult peripheral blood mononuclear cells (PBMCs) or cord blood (CB) cells before and after stimulation with heat-killed (HK) *L. monocytogenes* (Lm). These results are representative of ten individual PBMCs or cord blood samples. Error bars, mean \pm s.e.m.



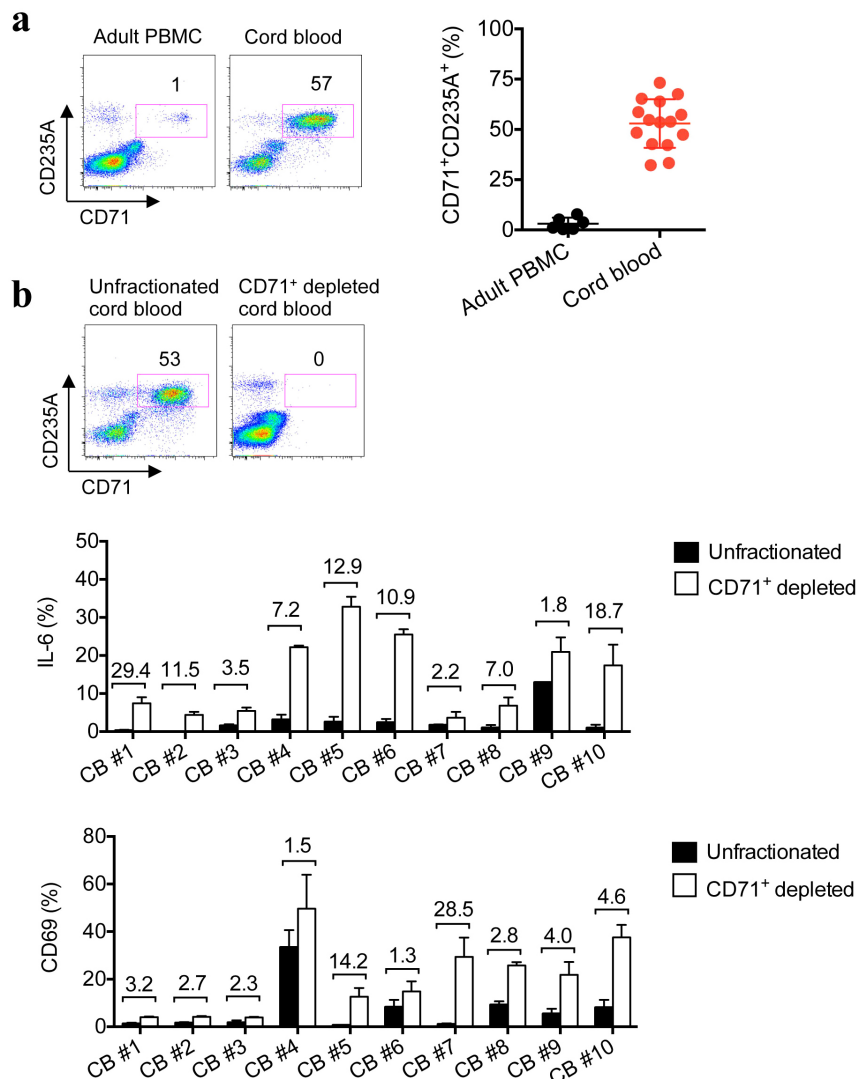
Extended Data Figure 3 | Neonatal splenocytes suppress the activation of adult cells in co-culture. **a**, TNF- α production by adult CD11c⁺ and B220⁺ responder cells after stimulation with heat-killed (HK) *L. monocytogenes* (Lm), and co-culture with the indicated ratio of splenocytes from 6-day-old neonatal mice. **b**, Representative plots and composite data illustrating CD69

and CD25 expression by adult CD8⁺ responder cells after stimulation with anti-CD3 antibody, and co-culture with the indicated ratios of neonatal splenocytes. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.



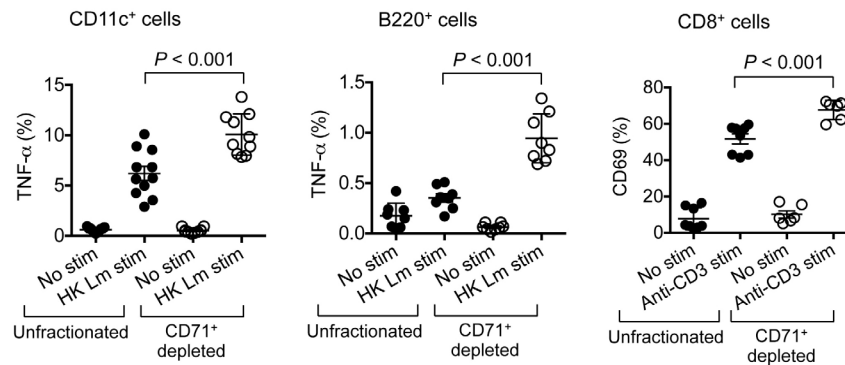
Extended Data Figure 4 | Arginase inhibition overrides the immunosuppressive properties of neonatal cells. TNF- α production by adult CD11b⁺ responder cells after stimulation with heat-killed (HK) *L. monocytogenes* (Lm) alone (black), or co-culture with neonatal splenocytes at a 1:1 ratio (blue), or additional supplementation (red) with arginase inhibitors (nor-NOHA or L-NOHA) (top), or 4-aminobenzoic acid hydrazide (a myeloperoxidase inhibitor), 4'-hydroxy-3'-methoxyacetophenone (an NADPH oxidase inhibitor), sodium diethyldithiocarbamate trihydrate

(a superoxide dismutase inhibitor), SB-431542 hydrate (a TGF- β receptor inhibitor), anti-TGF- β antibody (clone 1D11), Sn(IV) protoporphyrin IX dichloride (a haem oxidation inhibitor), acetylcysteine (a reactive oxygen species (ROS) scavenger), 1-methyl-D-tryptophan and 1-methyl-L-tryptophan (IDO inhibitors), and an anti-IL-10 receptor antibody (clone 1B1.3A) (bottom). Each point represents data from an individual mouse, and the data are representative of two independent experiments. Error bars, mean \pm s.e.m.



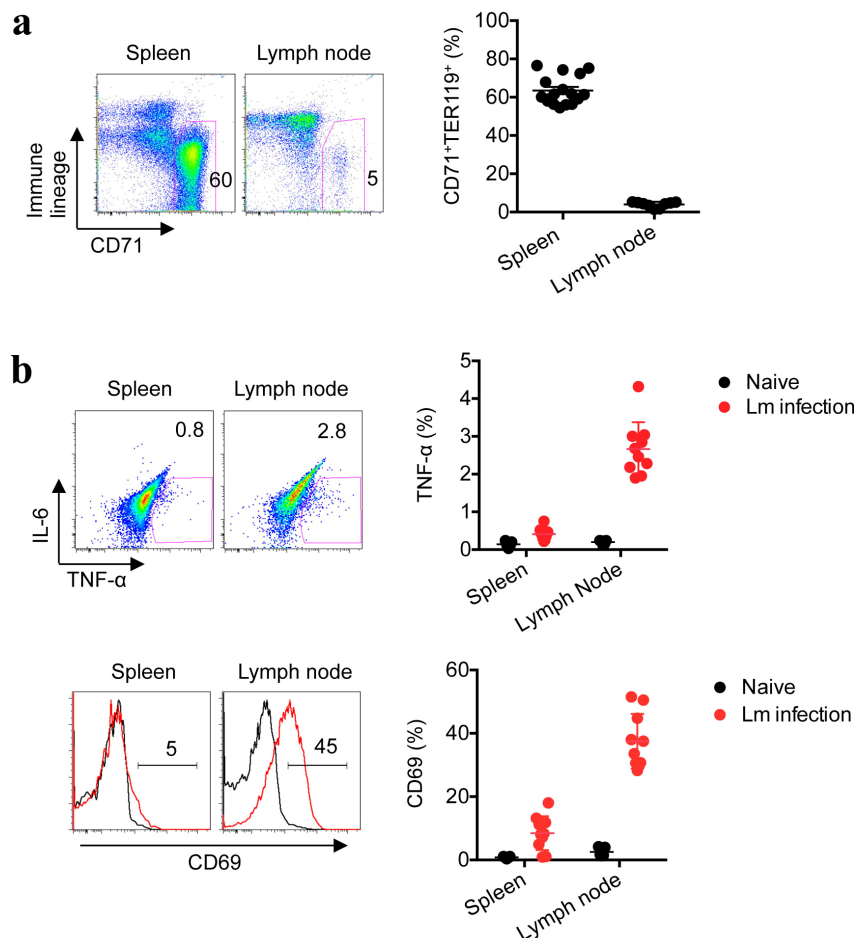
Extended Data Figure 5 | Human cord blood compared with adult peripheral blood mononuclear cells is enriched with CD71⁺CD235A⁺ erythroid cells, and the depletion of these erythroid cells unleashes the activation of the remaining neonatal immune cells. a, The proportion of CD71⁺CD235A⁺ cells among human adult PBMCs and cord blood. **b,** The proportion of CD71⁺CD235A⁺ cells among unfractionated and CD71⁺

cell-depleted cord blood (top), and the proportion of IL-6-producing cells among CD11b⁺ cells or the proportion of CD69⁺ cells among CD8⁺ cells in ten individual unfractionated and CD71⁺ cell-depleted cord blood specimens. The fold increase in CD71⁺ cell-depleted populations compared with unfractionated controls is shown above each pair of bars. Error bars, mean \pm s.e.m.



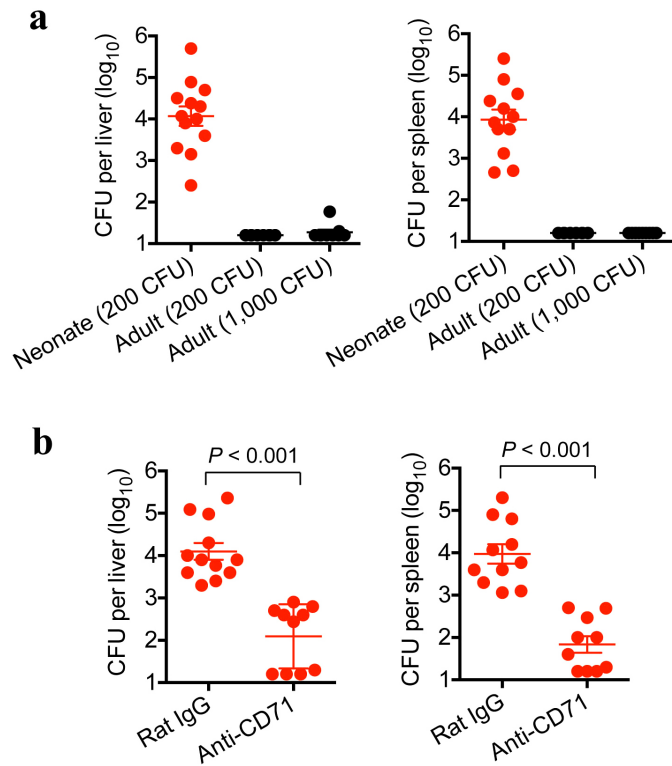
Extended Data Figure 6 | CD71⁺ cell depletion invigorates neonatal immune cell activation. TNF- α -producing CD11c⁺ or B220⁺ cells after stimulation with heat-killed (HK) *L. monocytogenes* (Lm), or CD69 expression by CD8⁺ T cells after stimulation with anti-CD3 antibody

among unfractionated or CD71⁺ cell-depleted splenocytes from 6-day-old neonatal mice. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.

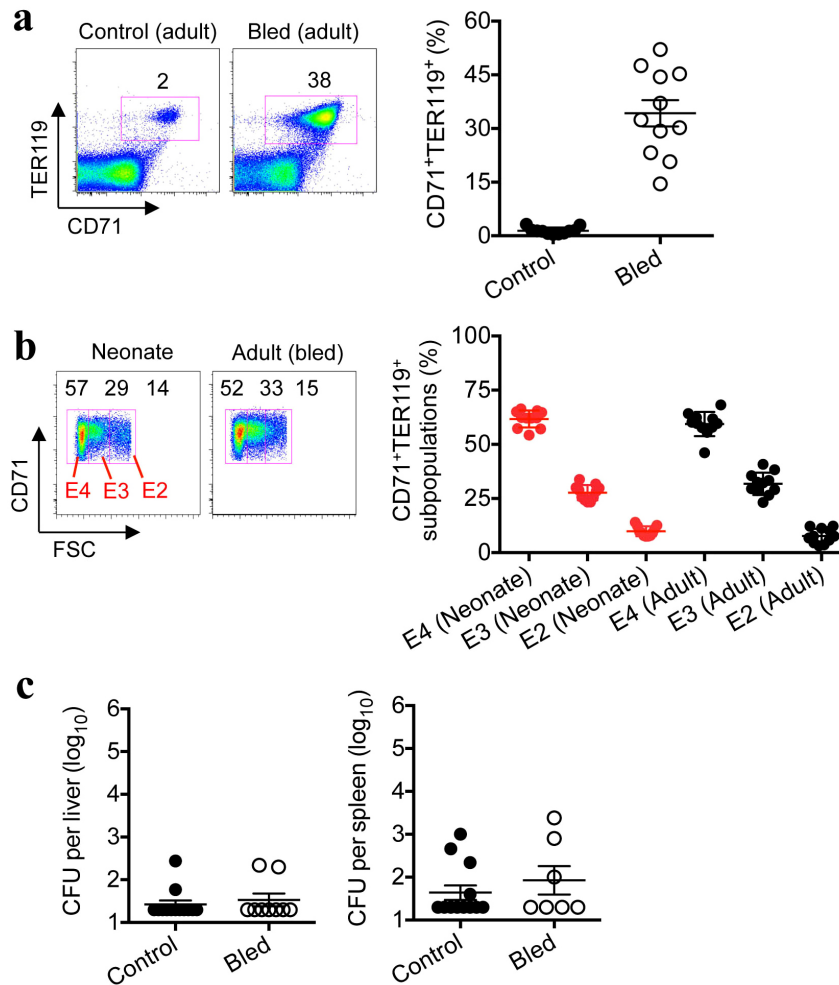


Extended Data Figure 7 | Infection-induced cell activation is enhanced in the neonatal lymph node, where immunosuppressive CD71⁺ erythroid cell numbers are diminished. **a**, The proportion of CD71⁺TER119⁺ cells among splenocytes and inguinal lymph node cells in 6-day-old neonatal mice. **b**, Representative plots and composite data comparing TNF- α production by

CD11b⁺ cells and CD69 expression by CD8⁺ cells 48 h after *L. monocytogenes* (Lm) infection (red line histogram) or in no infection controls (black line histogram) in neonatal mice. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.

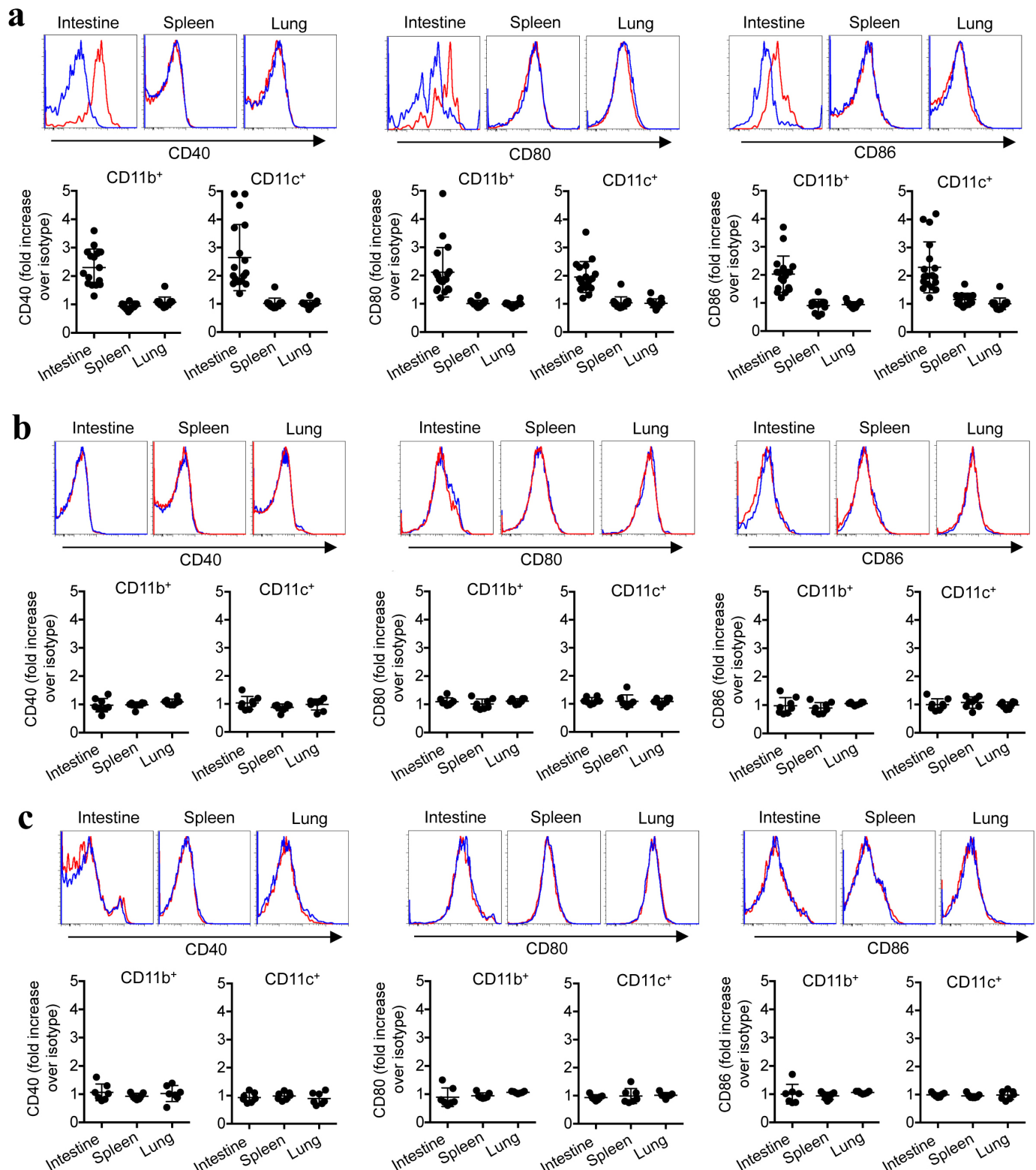


Extended Data Figure 8 | CD71⁺ cell depletion restores neonatal resistance to *E. coli* infection. **a**, Recoverable bacteria 48 h after infection with the indicated *E. coli* doses in 6-day-old (neonatal) or 8-week-old (adult) mice. **b**, Recoverable bacteria 48 h after *E. coli* infection (200 CFU) of anti-CD71-antibody-treated neonatal mice compared with isotype-control-antibody-treated neonatal mice. Anti-CD71 antibody or isotype control antibody (150 μ g) was administered on days 5 and 6 after birth, corresponding to 1 day before infection and the day of infection. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.



Extended Data Figure 9 | Phlebotomy-induced adult CD71⁺ erythroid cells have a similar subset distribution to neonatal cells but do not cause susceptibility to infection. **a**, The proportion of CD71⁺TER119⁺ splenocytes in adult control or phlebotomized (bled) mice 5 days after the initiation of daily phlebotomy for 3 consecutive days. **b**, The distribution of TER119⁺ erythroid cells based on CD71 expression and size (FSC, forward scatter) among

splenocytes from 6-day-old (neonatal) or phlebotomized 8-week-old (adult) mice. **c**, The number of recoverable bacteria after *L. monocytogenes* (Lm) infection (1,000 CFU). Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.



Extended Data Figure 10 | CD71⁺ erythroid cells selectively restrict immune cell activation in the intestine. **a**, Relative CD40, CD80 and CD86 expression by CD11b⁺ or CD11c⁺ cells recovered from the indicated tissues of 8-day-old (neonatal) mice treated with 150 μ g anti-CD71 antibody (red histograms gated on CD11b⁺ cells) or 150 μ g isotype control antibody (blue histograms gated on CD11b⁺ cells) on days 5 and 6. The mean fluorescence intensity for each parameter in the cells from CD71⁺ cell-depleted

mice was normalized to the levels in isotype-control-antibody-treated mice. **b**, CD40, CD80 and CD86 expression for the neonatal mice described in **a** after receiving antibiotic supplementation. **c**, CD40, CD80 and CD86 expression for the neonatal mice described in **a** after maintenance under gnotobiotic germ-free conditions. Each point represents data from an individual mouse, and the data are representative of three independent experiments. Error bars, mean \pm s.e.m.

EHMT1 controls brown adipose cell fate and thermogenesis through the PRDM16 complex

Haruya Ohno^{1*}, Kosaku Shinoda^{1*}, Kana Ohyama^{1*}, Louis Z. Sharp¹ & Shingo Kajimura¹

Brown adipose tissue (BAT) dissipates chemical energy in the form of heat as a defence against hypothermia and obesity. Current evidence indicates that brown adipocytes arise from *Myf5*⁺ dermatomal precursors through the action of PR domain containing protein 16 (PRDM16) transcriptional complex^{1,2}. However, the enzymatic component of the molecular switch that determines lineage specification of brown adipocytes remains unknown. Here we show that euchromatic histone-lysine *N*-methyltransferase 1 (EHMT1) is an essential BAT-enriched lysine methyltransferase in the PRDM16 transcriptional complex and controls brown adipose cell fate. Loss of EHMT1 in brown adipocytes causes a severe loss of brown fat characteristics and induces muscle differentiation *in vivo* through demethylation of histone 3 lysine 9 (H3K9me2 and 3) of the muscle-selective gene promoters. Conversely, EHMT1 expression positively regulates the BAT-selective thermogenic program by stabilizing the PRDM16 protein. Notably, adipose-specific deletion of EHMT1 leads to a marked reduction of BAT-mediated adaptive thermogenesis, obesity and systemic insulin resistance. These data indicate that EHMT1 is an essential enzymatic switch that controls brown adipose cell fate and energy homeostasis.

Obesity develops when energy intake chronically exceeds total energy expenditure. All anti-obesity medications currently approved by the FDA act to repress energy intake, either by suppressing appetite or by inhibiting intestinal fat absorption. However, because of their side effects including depression, oily bowel movements and steatorrhoea, there is an urgent need for alternative approaches. BAT is specialized to dissipate energy through uncoupling protein 1 (UCP1). Recent studies with ¹⁸F-fluoro-labelled 2-deoxy-glucose positron emission tomography (¹⁸FDG-PET) scanning demonstrated that adult humans have active BAT deposits^{3–6} and that the amount of BAT inversely correlates with adiposity and body mass index^{4,5}, indicating that it plays an important role in energy homeostasis in adult humans. Hence, a better understanding of the molecular control of BAT development may lead to an alternative approach to alter energy balance by increasing energy expenditure.

It has been reported that brown adipocytes in the interscapular and peri-renal BAT arise from *Engrailed-1*⁺ and *Myf5*⁺ dermatomal precursors^{1,7,8}. The PRDM16–C/EBP- β complex in the myogenic precursors activates the brown adipogenic gene program through inducing peroxisome proliferator-activated receptor (PPAR)- γ expression^{1,2,9}; however, the mechanism by which the PRDM16–C/EBP- β complex functions as a fate switch to control brown adipocyte versus myocyte lineage remains unexplored.

Previously we determined the essential domains of PRDM16 for converting myoblasts into brown adipocytes by generating two deletion mutants of PRDM16: a mutant lacking the PR-domain (Δ PR), a domain that shares high homology with methyltransferase SET domains^{10,11}, and a mutant lacking the zinc-finger domain-1 (Δ ZF-1) (Fig. 1a, top panel). Wild type (WT) and the Δ PR mutant, but not the Δ ZF-1 mutant, were able to convert myoblasts into brown adipocytes, suggesting that the ZF-1 domain is required². Consistent with the results, the PRDM16 complex purified from brown adipocytes expressing WT

and Δ PR, but not Δ ZF-1, had significant methyltransferase activities on H3 (Fig. 1a, bottom panel). Because this effect was independent of its SET domain, we searched for methyltransferases that were associated with differentiation-competent PRDM16 proteins (that is, WT and Δ PR), but not with differentiation-incompetent PRDM16 (Δ ZF-1). By using high-resolution liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS), we found EHMT1 as the only methyltransferase that was co-purified preferentially with the differentiation-competent PRDM16 complexes². EHMT1 has enzymatic activity on H3K9 mono- or di-Me¹². Notably, haploinsufficiency of the EHMT1

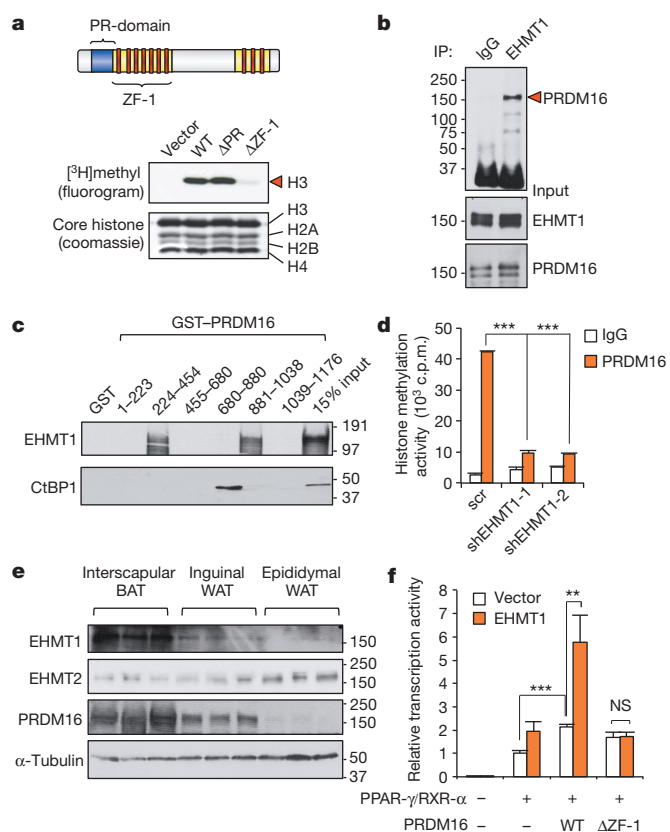


Figure 1 | Identification of EHMT1 in the PRDM16 transcriptional complex. **a**, Top: schematic illustration of PRDM16. Bottom: PRDM16 complex purified from brown adipocytes were subjected to *in vitro* histone methylation assay. **b**, Immunoprecipitation of EHMT1 protein followed by western blotting to detect PRDM16. Input is shown in lower panels. **c**, *In vitro* binding assay of ³⁵S-labelled EHMT1 or CtBP1 and purified PRDM16 fragments. **d**, Histone methylation assay of PRDM16 complex from brown adipocytes expressing indicated constructs ($n = 3$ or 4). **e**, Western blotting for indicated proteins in adipose tissues. **f**, Transcriptional activities of PRDM16 using a PPAR- γ -responsive luciferase reporter ($n = 3$). Error bars, s.e.m. ** $P < 0.01$, *** $P < 0.001$.

¹UCSF Diabetes Center, Department of Cell and Tissue Biology, University of California, San Francisco, 35 Medical Center Way, San Francisco, California 94143-0669, USA.

*These authors contributed equally to this work.

gene, because of 9q34.3 microdeletions or point mutations in humans¹³, is associated with clinical phenotypes including mental retardation. Importantly, 40–50% of patients with EHMT1 mutations develop obesity^{14,15}; however, the underlying mechanism remains completely unknown. Given the essential role of the PRDM16 complex for BAT development, we considered that EHMT1 is a key enzymatic component that controls the lineage specification and thermogenic function of BAT.

To test this hypothesis, we first confirmed the PRDM16–EHMT1 interaction by immunoprecipitation followed by western blotting in brown adipocytes (Fig. 1b and Supplementary Fig. 1). The purified ZF-1 (224–454) and ZF-2 (881–1038) domains of glutathione *S*-transferase (GST)–PRDM16 protein bound to the *in-vitro*-translated EHMT1 protein, whereas the 680–1038 region of PRDM16 bound to CtBP1 as previously reported¹⁶ (Fig. 1c and Supplementary Fig. 2). These results indicate that EHMT1 directly interacts with PRDM16. EHMT1 is the main methyltransferase of the PRDM16 complex in brown adipocytes, because the histone methyltransferase activity of the PRDM16 complex was largely lost when EHMT1 was depleted using two short hairpin RNAs (shRNAs) targeted to EHMT1 (Fig. 1d and Supplementary Fig. 3). Furthermore, expression of EHMT1 protein was highly enriched in BAT and in cultured brown adipocytes, correlating well with PRDM16 (Fig. 1e and Supplementary Fig. 4). In contrast, amounts of EHMT2 protein were higher in white adipose tissue (WAT) than in BAT. To test if EHMT1 modulates the PRDM16 transcriptional activity, we

performed luciferase assays using a luciferase reporter gene containing PPAR- γ binding sites¹. As shown in Fig. 1f, co-expression of EHMT1 and PRDM16 synergistically increased the reporter gene activity, whereas this induction was completely lost when the Δ ZF-1 mutant was expressed. These data indicate that EHMT1 forms a transcriptional complex with PRDM16 and regulates its activity through direct interaction.

Next, we investigated the genetic requirement for EHMT1 in BAT development *in vivo*. Because a whole-body knockout of the *Ehmt1* gene causes embryonic lethality before the emergence of brown adipocytes¹², the *Ehmt1* gene was deleted in brown adipocyte precursors by crossing *Ehmt1*^{fllox/fllox} mice¹⁷ with *Myf5-Cre* mice¹. As shown in Fig. 2a–c, the interscapular BAT of *Ehmt1*^{myf5} knockout mice was substantially smaller than in WT mice at postnatal stage (P)1. Haematoxylin and eosin staining showed that brown adipocytes in *Ehmt1*^{myf5} knockout mice were significantly smaller and contained fewer lipids than in WT mice, whereas other tissues near the BAT including skin seemed normal (Fig. 2b and Supplementary Fig. 5). Similar results were observed in embryos at embryonic day (E)18.5 (Supplementary Fig. 6). Subsequently, we analysed the global gene expression of BAT from the WT and *Ehmt1*^{myf5} knockout embryos by RNA-sequencing. The following gene ontology analysis found that the gene expression pattern in the *Ehmt1*^{myf5} knockout BAT showed a skeletal-muscle phenotype: that is, a broad activation of the skeletal muscle-selective

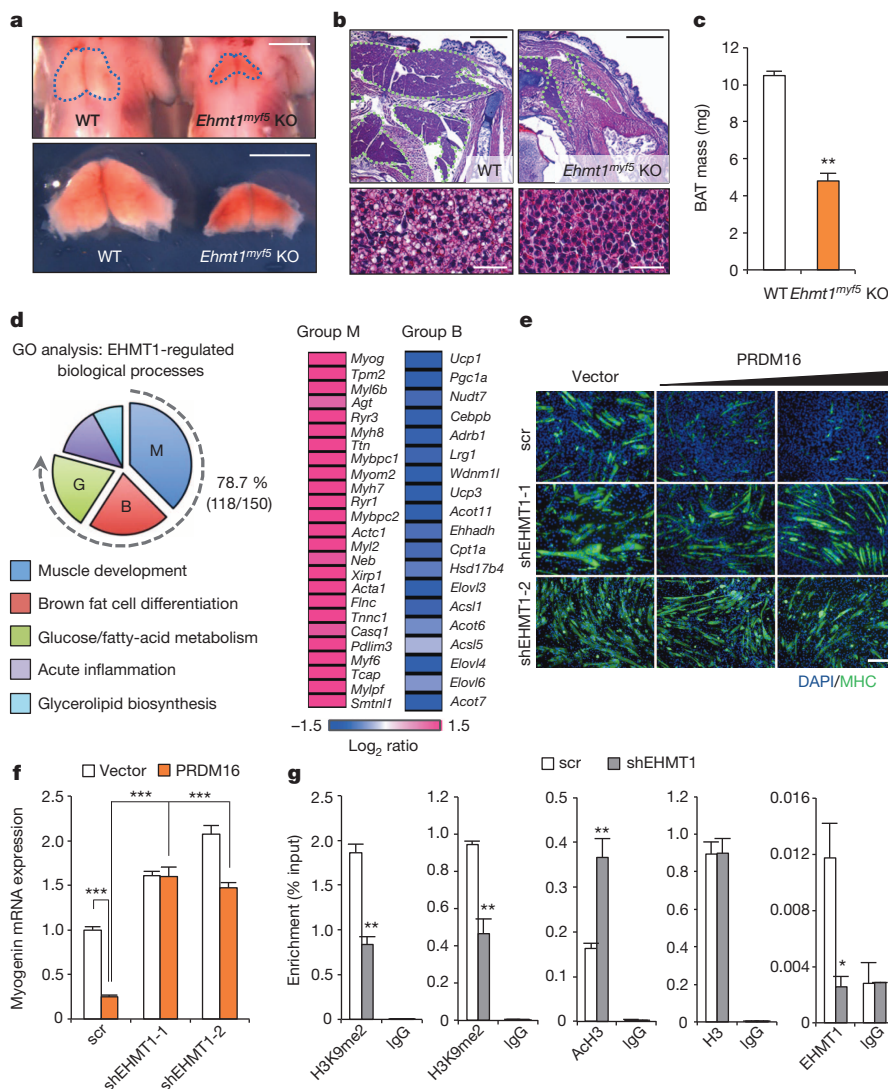


Figure 2 | EHMT1 is required for BAT versus muscle lineage specification. **a**, Morphology of BAT from WT and *Ehmt1*^{myf5} knockout embryos at P1. Scale bar, 2.5 mm. KO, knockout. **b**, Haematoxylin and eosin staining of WT and *Ehmt1*^{myf5} knockout (KO) BAT. Scale bar, 600 μ m. Bottom: high-magnification images. Scale bar, 30 μ m. **c**, BAT weight from WT ($n = 14$) and knockout embryos ($n = 8$). **d**, Gene ontology analyses of RNA-sequencing data. The log₂-fold changes in the expression of skeletal muscle (group M) and BAT (group B) genes are shown. **e**, Immunocytochemistry for MHC in C2C12 cells expressing indicated constructs under pro-myogenic culture conditions. Scale bar, 200 μ m. **f**, Myogenin mRNA expression in **e** ($n = 3$). **g**, Chromatin immunoprecipitation assays using indicated antibodies ($n = 3$). Error bars, s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

genes, and a broad reduction of the BAT-selective genes. Strikingly, 78.7% of the differentially expressed genes (118 out of 150 genes) between WT and knockout mice were stratified into categories of skeletal muscle development, BAT development and BAT function (glucose/fatty-acid metabolism). Specifically, 77.5% of the ectopically activated genes in the knockout BAT were related to skeletal muscle development, including myogenin and myosin heavy chains. On the other hand, 80.0% of the reduced genes in the knockout BAT were involved in BAT development and fatty-acid/glucose metabolism, including *Ucp1*, *Pgc1a*, *Cebpb*, *Cpt1a* and *Elovl3* (Fig. 2d and Supplementary Fig. 7). These results indicate that EHMT1 is absolutely required for the cell-fate specification between BAT versus muscle.

To investigate the mechanisms by which EHMT1 determines BAT lineage, retroviruses expressing a scrambled control RNA (scr) or shRNAs targeting EHMT1 (shEHMT1-1 and -2) were transduced into C2C12 myoblasts together with PRDM16 (Supplementary Fig. 8a). As shown in Fig. 2e (upper panels), PRDM16 expression powerfully blocked myogenic differentiation in a dose-dependent fashion, as shown by immunohistochemistry using a pan-skeletal myosin heavy chain (MHC) antibody. In contrast, EHMT1 depletion significantly impaired the PRDM16-mediated repression on myogenesis (Fig. 2e, lower panels). Gene expression analysis showed that the repression on muscle-selective genes such as myogenin was near completely abolished when EHMT1 was depleted (Fig. 2f and Supplementary Fig. 8b). This repressive effect was mediated through the methyltransferase activity of EHMT1, because ectopic expression of the EHMT1 mutant (N1198L;H1199E) that lacks methyltransferase activity¹⁸ significantly blunted the PRDM16-mediated repression on myogenesis (Supplementary Fig. 9a). Additionally, two chemical inhibitors of EHMT1/2, BIX-01294 and UNC0638, blocked the repressive effects of PRDM16 (Supplementary Fig. 9b, c). BIX-01294 treatment in brown adipocytes also significantly reduced the expression of BAT-selective genes (Supplementary Fig. 9d). Consistent with these data, chromatin immunoprecipitation assays found that EHMT1 depletion robustly reduced amounts of H3K9me2 and me3 at the proximal region of the myogenin gene promoter on which EHMT1 was recruited (Fig. 2g). On the contrary, the amounts of H3K9/14ac were significantly increased by EHMT1 depletion without any effect on total H3 amounts. Similar changes were observed at the promoter regions of other muscle-selective genes including *Act1*, *Ryr1* and *Myh1*, where EHMT1 was recruited (Supplementary Fig. 10). Conversely, under pro-adipogenic culture conditions, knockdown of EHMT1 largely blocked the PRDM16-induced brown adipogenesis in C2C12 cells (Supplementary Fig. 11). Together, these results indicate that EHMT1 determines BAT versus muscle cell lineage through PRDM16 by controlling H3K9 methylation status of the muscle-selective gene promoters.

To investigate the role of EHMT1 in BAT thermogenesis, EHMT1 was depleted in immortalized brown adipocytes by retrovirus-mediated shRNA knockdown (Supplementary Fig. 12a, b). Total and uncoupled (oligomycin-insensitive) oxygen consumption rate in the EHMT1-depleted brown adipocytes was significantly reduced both at the basal and cyclic AMP (cAMP)-stimulated states (Fig. 3a). Conversely, EHMT1 overexpression significantly increased messenger RNA (mRNA) amounts of BAT-selective thermogenic genes, including *Ucp1*, *Pgc1a* and *Dio2* (Fig. 3b), and oxygen consumption rate (Supplementary Fig. 12c). To test further if this EHMT1 action requires PRDM16, EHMT1 was ectopically introduced into mouse embryonic fibroblasts that did not express endogenous PRDM16. As shown in Fig. 3c, mouse embryonic fibroblasts expressing PRDM16 and C/EBP- β uniformly differentiated into lipid-containing adipocytes as previously reported². Although EHMT1 alone did not stimulate brown adipogenesis, the combination of EHMT1 with PRDM16 and C/EBP- β synergistically increased mRNA amounts of the BAT-selective genes, including *Ucp1*, *Cidea*, *Cox7a* and *Cox8b* (Fig. 3d). These data indicate that EHMT1 positively regulates the BAT-selective thermogenic gene program through PRDM16.

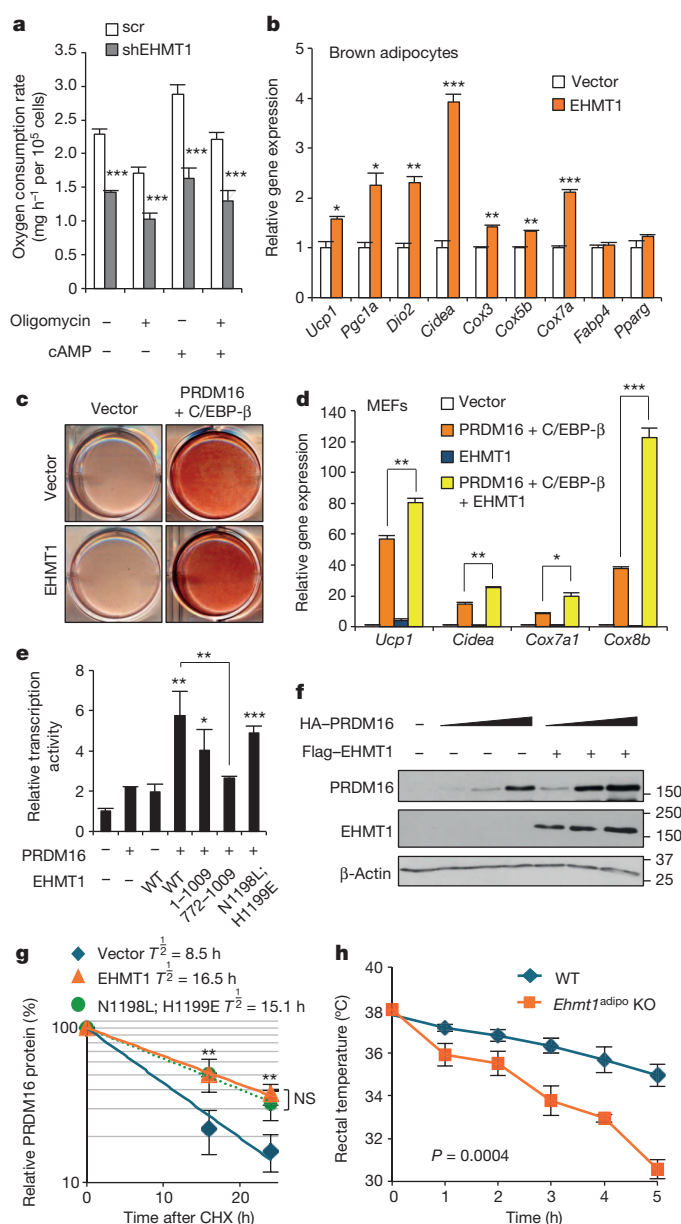


Figure 3 | EHMT1 controls BAT thermogenesis through stabilizing PRDM16 protein. **a**, Cellular respiration in brown adipocytes expressing indicated constructs ($n = 6$). **b**, BAT-selective gene expression in brown adipocytes expressing indicated constructs ($n = 3$). **c**, Oil-Red-O staining of mouse embryonic fibroblasts expressing indicated constructs under pro-adipogenic culture conditions. **d**, BAT-selective gene expression in C2C12 cells ($n = 3$). **e**, Effects of EHMT1 mutants on PRDM16 transcriptional activities ($n = 3$). **f**, Amounts of PRDM16 protein in COS7 cells expressing indicated constructs. **g**, Regression analysis of the PRDM16 protein stability ($n = 3$). **h**, Changes in rectal temperature during a cold challenge ($n = 4$ or 5). Error bars, s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

To determine which domains of EHMT1 are required for the induction of PRDM16 transcriptional activity, we tested a series of EHMT1 mutants for their ability to interact and co-localize with PRDM16 (Supplementary Figs 13 and 14). A deletion mutant of EHMT1 (772–1009), which failed to interact with PRDM16, was unable to increase PRDM16 reporter gene activity (Fig. 3e). EHMT1 with mutations in the SET domain (N1198L; H1199E) had no methyltransferase activity, but was still able to bind to and activate PRDM16. Even a mutant that lacked the SET-domain altogether (1–1009) interacted with and activated PRDM16. Thus an interaction between EHMT1 and PRDM16 seems to be required to activate PRDM16 transcriptional activity. Notably,

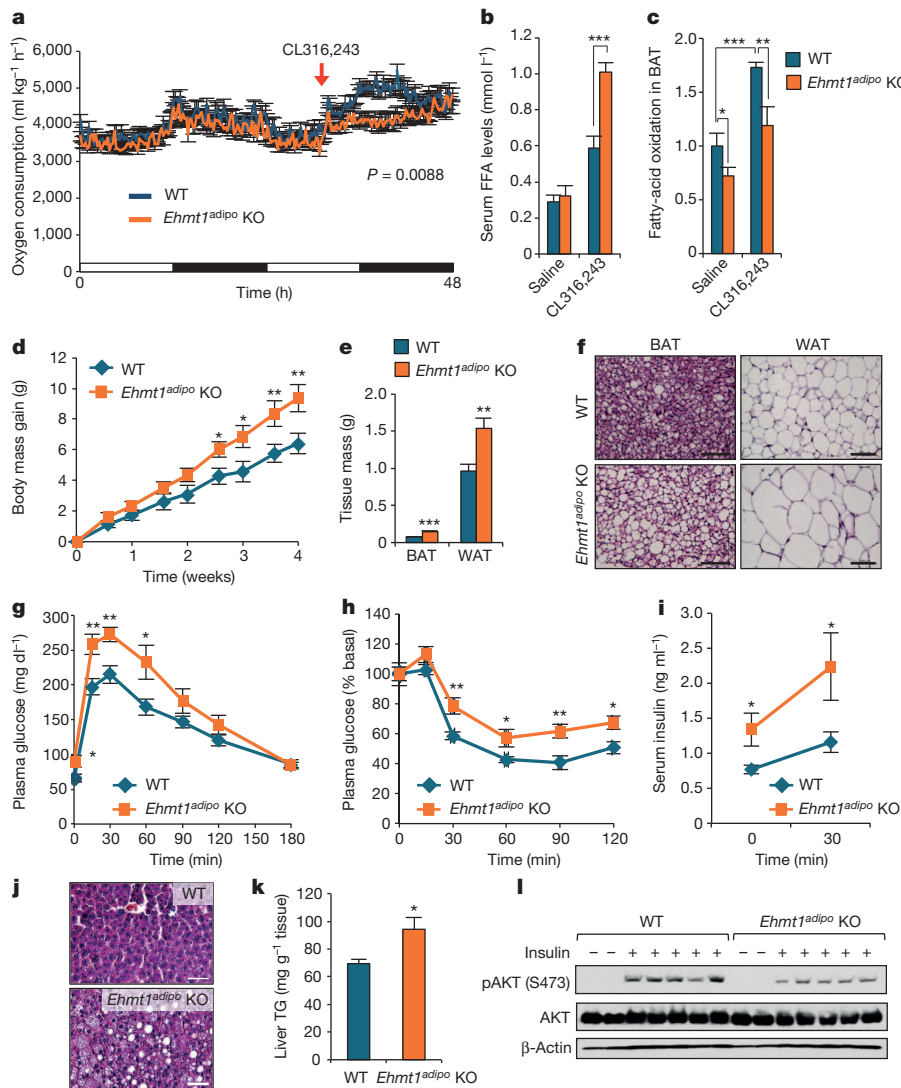


Figure 4 | EHMT1 deficiency in BAT causes obesity and insulin resistance. **a**, Oxygen consumption rate of WT and *Ehmt1^{adipo}* knockout mice treated with CL316,243 (0.5 mg kg⁻¹) at thermoneutrality ($n = 6$). **b**, Amounts of serum FFA in mice treated with saline or CL316,243. **c**, Fatty-acid oxidation in BAT ($n = 6-10$). **d**, Body mass change under a high-fat diet at thermoneutrality ($n = 16$). **e**, Adipose tissue mass after 4-week high-fat diet ($n = 16$). **f**, Haematoxylin and eosin staining of adipose tissues. Scale bar, 100 μ m. **g**, Glucose tolerance test in 9-week high-fat diet-fed mice ($n = 9$). **h**, Insulin tolerance test in 10-week high-fat diet-fed mice ($n = 9$). **i**, Amounts of serum insulin at the fasted and glucose-stimulated states ($n = 9$). **j**, Haematoxylin and eosin staining of liver in **d**. Scale bar, 50 μ m. **k**, Amounts of liver triglyceride in **j** ($n = 9$). **l**, Hepatic insulin signalling as assessed by phosphorylated (S473) and total Akt amounts. Error bars, s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

expression of EHMT1 robustly increased the amount of PRDM16 protein, independently of its mRNA expression (Fig. 3f and Supplementary Fig. 15). This effect was due to changes in the rate of protein degradation, because cycloheximide chase experiments showed that expression of EHMT1 extended the half-life of PRDM16 protein from 8.5 to 16.5 h. The N1198L;H1199E mutant also extended the half-life of PRDM16 protein as potently as the WT form (Fig. 3g). PRDM16 protein accumulation was induced only by the EHMT1 mutants that bind to PRDM16 (Supplementary Fig. 16). EHMT1 regulates endogenous PRDM16 protein amounts *in vivo* (Extended Data 1). EHMT1 protein stability was not affected by PRDM16 (Supplementary Fig. 17). These results collectively suggest that EHMT1 has dual functions: that is, repressive effects on the muscle-selective gene program through its methyltransferase activity, and activation of the BAT-selective gene program through stabilization of PRDM16 protein through direct association.

Next, we examined the requirement for EHMT1 in adaptive thermogenesis *in vivo*. To exclude potential defects in the skeletal muscle of *Ehmt1^{myf5}* knockout mice, we generated adipose tissue-specific *Ehmt1* knockout mice (*Ehmt1^{adipo}* knockout) using *Adiponectin-Cre* mice¹⁹. Of note, 62.3% of the differentially expressed muscle/BAT-selective genes in the *Ehmt1^{myf5}* mice were similarly dysregulated in the *Ehmt1^{adipo}* knockout mice (Extended Data 2). Although *Adiponectin-Cre* is expressed both in BAT and WAT, expression of EHMT1 is highly enriched in BAT compared with WAT (Fig. 1e). Furthermore, lipolysis capacity in the WAT of *Ehmt1^{adipo}* knockout mice was indistinguishable

from WT mice (Supplementary Fig. 18). EHMT1 is required for beige/brite cell development (Extended Data 3). Hence, the *Ehmt1^{adipo}* knockout mice allow us to examine the role of EHMT1 in BAT/beige fat-mediated thermogenesis *in vivo*. As shown in Fig. 3h, rectal temperature of *Ehmt1^{adipo}* knockout mice strikingly dropped within 1 h after a cold challenge to 4 °C, whereas that of control mice remained constant. Expression of BAT-selective genes in skeletal muscle²⁰ was not altered in *Ehmt1^{adipo}* knockout mice (Supplementary Fig. 19). We subsequently measured oxygen consumption rate at thermoneutrality (29–30 °C)²¹ in response to an activation of the β 3-adrenoceptor pathway. As shown in Fig. 4a, the oxygen consumption rate of WT mice was significantly increased after administering CL316,243 whereas this induction was completely lost in knockout mice. The impaired thermogenesis in knockout mice was accompanied by higher serum amounts of free fatty acids (FFAs) (Fig. 4b). This is consistent with previous findings that BAT serves as a major sink of FFAs for heat generation²², and that reduced fatty-acid oxidation in BAT leads to an increase in amounts of serum FFA²³. Indeed, fatty-acid oxidation capacity in the knockout BAT was significantly lower than in WT mice at the basal state and after administering CL316,243 (Fig. 4c). Additionally, fatty-acid uptake in the knockout BAT was reduced (Supplementary Fig. 20). These results indicate that EHMT1 is absolutely required for BAT-mediated adaptive thermogenesis and fatty-acid metabolism *in vivo*.

Lastly, we tested whether EHMT1 deficiency in BAT affects the propensity for weight gain in response to an obesogenic diet at thermoneutrality (29–30 °C), because an obesity phenotype in *Ucp1* knockout

mice was observed only at thermoneutrality²⁴. As shown in Fig. 4d and Supplementary Fig. 21, *Ehmt1^{adipo}* knockout mice gained significantly more body mass than WT mice without any change in food intake (Supplementary Table 1). Knockout mice had higher amounts of epididymal WAT and interscapular BAT that contained substantially larger lipid droplets than WT mice (Fig. 4e, f). A glucose tolerance test found that knockout mice showed significantly higher blood glucose concentrations than WT mice (Fig. 4g). Similarly, knockout mice showed impaired responses to insulin during an insulin tolerance test (Fig. 4h) and higher amounts of serum insulin at the fasted and glucose-stimulated states (Fig. 4i). Knockout mice showed an insulin-resistance phenotype even at ambient temperature, whereas no statistically significant difference was observed in body mass (Supplementary Fig. 22). Notably, the liver from knockout mice contained higher amounts of lipids and triglyceride (Fig. 4j, k) and showed impaired insulin signalling as assessed by phosphorylation of Akt in response to insulin (Fig. 4l and Supplementary Fig. 23). Together, these results indicate that EHMT1 deficiency in BAT leads to obesity, systemic insulin resistance and hepatic steatosis under a high-fat diet.

In conclusion, we have identified EHMT1 as an essential BAT-enriched methyltransferase that controls brown adipose cell fate, adaptive thermogenesis and glucose homeostasis *in vivo*. Although presence of BAT in adult humans is now widely appreciated, no mutation that causes defects in human BAT development and thermogenesis had been described except polymorphisms in *UCP1* and β 3-adrenoceptor genes²⁵. Delineating the causal link between EHMT1 mutations and BAT thermogenesis will provide a new perspective in understanding the molecular control of energy homeostasis through the epigenetic pathways, which may lead to effective therapeutic interventions for obesity and metabolic diseases.

METHODS SUMMARY

Animals. All animal experiments were performed according to procedures approved by University of California, San Francisco's Institutional Animal Care and Use Committee. *Ehmt1^{lox17}* and *Adiponectin-Cre* mice¹⁹ were provided by A. Tarakhovsky and E. D. Rosen. For metabolic studies, male mice in Bl6 background were fed with a high-fat diet for 4 weeks at thermoneutrality and ambient temperature.

Bioinformatics. RNA-sequencing libraries were constructed at the University of California, San Francisco Genomic Core Facility. Gene ontology enrichment analyses were performed on the differentially expressed genes ($P < 0.05$, the delta-method-based hypothesis test) using RefSeq as the background data set. The accession number for the data is E-MTAB-1704.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 February; accepted 13 September 2013.

Published online 6 November 2013.

- Seale, P. *et al.* PRDM16 controls a brown fat/skeletal muscle switch. *Nature* **454**, 961–967 (2008).
- Kajimura, S. *et al.* Initiation of myoblast to brown fat switch by a PRDM16–C/EBP- β transcriptional complex. *Nature* **460**, 1154–1158 (2009).
- Nedergaard, J., Bengtsson, T. & Cannon, B. Unexpected evidence for active brown adipose tissue in adult humans. *Am. J. Physiol.* **293**, E444–E452 (2007).
- van Marken Lichtenbelt, W. D. *et al.* Cold-activated brown adipose tissue in healthy men. *N. Engl. J. Med.* **360**, 1500–1508 (2009).
- Saito, M. *et al.* High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. *Diabetes* **58**, 1526–1531 (2009).

- Virtanen, K. A. *et al.* Functional brown adipose tissue in healthy adults. *N. Engl. J. Med.* **360**, 1518–1525 (2009).
- Atit, R. *et al.* β -catenin activation is necessary and sufficient to specify the dorsal dermal fate in the mouse. *Dev. Biol.* **296**, 164–176 (2006).
- Timmons, J. A. *et al.* Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages. *Proc. Natl Acad. Sci. USA* **104**, 4401–4406 (2007).
- Kajimura, S., Seale, P. & Spiegelman, B. M. Transcriptional control of brown fat development. *Cell Metab.* **11**, 257–262 (2010).
- Shing, D. C. *et al.* Overexpression of sPRDM16 coupled with loss of p53 induces myeloid leukemias in mice. *J. Clin. Invest.* **117**, 3696–3707 (2007).
- Pinheiro, I. *et al.* Prdm3 and Prdm16 are H3K9me1 methyltransferases required for mammalian heterochromatin integrity. *Cell* **150**, 948–960 (2012).
- Tachibana, M. *et al.* Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3–K9. *Genes Dev.* **19**, 815–826 (2005).
- Kleefstra, T. *et al.* Disruption of the gene euchromatin histone methyl transferase 1 (Eu-HMTase1) is associated with the 9q34 subtelomeric deletion syndrome. *J. Med. Genet.* **42**, 299–306 (2005).
- Cormier-Daire, V. *et al.* Cryptic terminal deletion of chromosome 9q34: a novel cause of syndromic obesity in childhood? *J. Med. Genet.* **40**, 300–303 (2003).
- Willemsen, M. H. *et al.* Update on Kleefstra syndrome. *Mol. Syndromol.* **2**, 202–212 (2012).
- Kajimura, S. *et al.* Regulation of the brown and white fat gene programs through a PRDM16/CtBP transcriptional complex. *Genes Dev.* **22**, 1397–1409 (2008).
- Schaefer, A. *et al.* Control of cognition and adaptive behavior by the GLP/G9a epigenetic suppressor complex. *Neuron* **64**, 678–691 (2009).
- Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H. & Shinkai, Y. G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. *EMBO J.* **27**, 2681–2690 (2008).
- Eguchi, J. *et al.* Transcriptional control of adipose lipid handling by IRF4. *Cell Metab.* **13**, 249–259 (2011).
- Almind, K., Manieri, M., Sivitz, W. I., Cinti, S. & Kahn, C. R. Ectopic brown adipose tissue in muscle provides a mechanism for differences in risk of metabolic syndrome in mice. *Proc. Natl Acad. Sci. USA* **104**, 2366–2371 (2007).
- Cannon, B. & Nedergaard, J. Nonshivering thermogenesis and its adequate measurement in metabolic studies. *J. Exp. Biol.* **214**, 242–253 (2011).
- Ouellet, V. *et al.* Outdoor temperature, age, sex, body mass index, and diabetic status determine the prevalence, mass, and glucose-uptake activity of 18F-FDG-detected BAT in humans. *J. Clin. Endocrinol. Metab.* **96**, 192–199 (2012).
- Wu, Q. *et al.* Fatty acid transport protein 1 is required for nonshivering thermogenesis in brown adipose tissue. *Diabetes* **55**, 3229–3237 (2006).
- Feldmann, H. M., Golozoubova, V., Cannon, B. & Nedergaard, J. UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. *Cell Metab.* **9**, 203–209 (2009).
- Yoneshiro, T. *et al.* Impact of UCP1 and beta3AR gene polymorphisms on age-related changes in brown adipose tissue and adiposity in humans. *Int. J. Obes.* **37**, 993–998 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to A. Tarakhovsky, E. D. Rosen, Y. Shinkai and E. Hara for providing mice and plasmids. We thank our colleagues in the University of California, San Francisco, including Y. Qiu, A. Chawla, C. Paillart, S. Koliwad, M. Robblee, D. Scheel, S. Ohata, L. Mera, D. Lowe, S. Sonne, S. Keylin, I. Luijten, H. Hong and E. Tomoda for their assistance. This work was supported by grants from the National Institutes of Health (DK087853 and DK97441) to S.K. We acknowledge supports from the DERC center grant (DK63720), University of California, San Francisco Program for Breakthrough Biomedical Research program, the Pew Charitable Trust, and PRESTO from the Japan Science and Technology Agency to S.K. H.O. is supported by the Manpei Suzuki Diabetes Foundation. K.S. is supported by a fellowship from the Japan Society for the Promotion of Science.

Author Contributions S.K. and H.O. conceived and designed the experiments. All authors performed the experiments and analysed the data. S.K. and H.O. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.K. (skajimura@diabetes.ucsf.edu).

METHODS

Animals. All animal experiments were performed according to procedures approved by University of California, San Francisco's Institutional Animal Care and Use Committee for animal care and handling. *Ehmt1^{flox}* mice and *Adiponectin^{Cre/+}* mice were provided by A. Tarakhovskiy and E. D. Rosen. *Myf5^{Cre/+}* mice were obtained from the Jackson Laboratory²⁶. To analyse embryonic BAT development, *Ehmt1^{flox/flox}* or *Myf5^{Cre/+}; Ehmt1^{flox/flox}* embryos at E18.5 or newborn mice at P1 were collected and fixed in 4% paraformaldehyde for histological analyses. The presumptive BAT depots in the interscapular region were micro-dissected for histological and RNA expression analyses. For cold exposure experiments, male *Ehmt1^{adipo}* knockout mice (*Adipo-Cre^{+/-}; Ehmt1^{flox/flox}*) and body-weight-matched control mice (*Ehmt1^{flox/flox}*) at 10 weeks of age were single-caged and exposed to 4 °C for 5 h. Rectal temperatures were monitored every hour using a TH-5 thermometer (Physitemp).

Metabolic studies. Whole-body energy expenditure of *Ehmt1^{adipo}* knockout mice or control mice matched for body mass at 14 weeks of age was measured at thermoneutrality (29–30 °C) using a Comprehensive Lab Animal Monitoring System (Columbus Instruments). The mice were injected intraperitoneally with a β 3-adrenergic receptor-specific agonist CL316,243 at a dose of 0.5 mg kg⁻¹. For diet-induced obesity studies, male mice at 6–7 weeks of age were fed a high-fat diet (D12492, Research Diet) for 4 weeks at thermoneutrality. At the end of the experiments, serum samples were collected. Amounts of serum insulin (Millipore), triglyceride (Thermo) and FFA (Wako) were measured using commercially available kits. For glucose tolerance test experiments, male mice were fed a high-fat diet for 9 weeks. After an overnight fast, the mice were injected intraperitoneally with glucose (1 g kg⁻¹). For insulin tolerance test experiments, male mice under a high-fat diet for 10 weeks were used. After an overnight fast, the mice were injected intraperitoneally with insulin (0.75 U kg⁻¹). Blood samples were collected at indicated time points and amounts of glucose were measured using blood glucose test strips (Abbott). To measure liver triglyceride contents, the liver tissue (25 mg) was homogenized in 1.25 ml of Folch solution (chloroform/methanol, 2:1, v/v). Subsequently, equal amounts (0.4 ml) of chloroform and water were added to the lysate. After centrifugation at 735g for 3 min, the chloroform phase was collected and dried. The pellet was dissolved in isopropanol. Amounts of triglycerides were determined by an Infinity Triglycerides kit (Thermo).

Fatty-acid oxidation assay. WT and *Ehmt1^{adipo}* knockout mice at 11 weeks old were intraperitoneally injected with saline or CL316,243. Five hours after the injection, the interscapular BAT depots were isolated. Fatty-acid oxidation assay was performed according to the protocol described by Mao *et al.*²⁷ Briefly, the adipose tissues were minced to small pieces and incubated with DMEM supplemented with 1 mM pyruvate, 1% FFA-free BSA and 0.5 mM oleate. [¹⁴C]oleic acid at 1 μ Ci μ l⁻¹ was added for 2 h at 37 °C. After adding 70% perchloric acid into each well, CO₂ was captured by Whatman paper soaked in 3 M NaOH solution for 1 h. [¹⁴C] radioactivity was measured by liquid scintillation counter and normalized to tissue mass. To assess fatty-acid uptake, BAT (approximately 100 mg) was isolated from WT and *Ehmt1^{adipo}* knockout mice and incubated in DMEM containing oleic acid (250 μ M, Nu-Chek Prep) supplemented with [¹⁴C]oleic acid (0.25 μ Ci μ l⁻¹) and 10% FBS for 15 min. [¹⁴C] radioactivity in the BAT explants was measured by liquid scintillation counter and normalized to the total protein content.

In vivo insulin stimulation assay. Mice were anaesthetized with Tribromoethanol (Avertin). Insulin (5 U) was injected into the inferior vena cavae. Livers were removed 2 min after the injection and lysed in lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10% (w/v) glycerol, 100 mM NaF, 10 mM EGTA, 1 mM Na₃VO₄, 1% (w/v) Triton X-100, 5 μ M ZnCl₂, 2 mM), with protease inhibitor cocktail (cOmplete, Roche). The lysates were isolated and separated by SDS–polyacrylamide gel electrophoresis (SDS–PAGE). Akt (Pan) antibody (Cell Signaling) and Phospho-Akt (Ser473) Antibody (Cell Signaling) were used for western blotting.

Fat lipolysis assay. Epididymal fat pads were collected and digested in a digestion buffer (121 mM NaCl, 4.9 mM KCl, 1.2 mM MgSO₄, 0.33 mM CaCl₂, 12 mM HEPES) containing collagenase D (1.5 U ml⁻¹), dipase II (2.4 U ml⁻¹), 3 mM glucose and fatty-acid-free 1% BSA (Akron Biotech). After digestion for 1 h at 37 °C with gentle shaking, adipocytes were filtrated through nylon mesh and centrifuged at 186g for 5 min. Floating adipocytes were collected and incubated in DMEM containing 10% FBS in the presence or absence of isoproterenol (1 μ M) for 1.5 h at 37 °C. Glycerol release into the media was determined using a free glycerol reagent (Sigma). Amounts of glycerol were normalized to the total protein content of the primary adipocytes by using a Pierce BCA Protein Assay reagent (Thermo Scientific).

Cell culture. Immortalized brown fat cells were isolated from the interscapular BAT of WT mice at P1–P3. Mouse embryonic fibroblasts have been described previously². HEK293 cells and C2C12 cells were obtained from the American Type Culture Collection. Adipocyte differentiation in C2C12 cells was induced by treating confluent cells with DMEM containing 10% FBS, 0.5 mM isobutylmethylxanthine, 125 μ M indomethacin, 2 μ g ml⁻¹ dexamethasone, 850 nM insulin, 1 nM

T3 and 0.5 μ M rosiglitazone. Two days after induction, cells were switched to the maintenance medium containing 10% FBS, 850 nM insulin, 1 nM T3 and 0.5 μ M rosiglitazone. For cAMP treatment, cells were incubated with 10 μ M forskolin for 4 h. Myocyte differentiation in C2C12 myoblasts was induced by treating cells in DMEM containing 2% horse serum. For beige cell differentiation in culture, the stromal vascular (SV) fraction was isolated from *Ehmt1^{flox/flox}* mice and plated in collagen-coated plates (BD Biosciences). Cells were differentiated in the absence or presence of rosiglitazone at 0.5 μ M according to the previous paper²⁸.

DNA constructs and viruses production. Deletion mutants of Flag-tagged PRDM16 and GST-fused PRDM16 fragments (1–223, 224–454, 455–680, 680–880, 881–1038 and 1039–1176) were described previously¹⁶. EHMT1 expression constructs were gifts from Y. Shinkai¹⁸ and E. Hara²⁹. EHMT1 was cloned to pMSCV-puro vector for retroviral expression. The sequences used for retroviral shRNA expression vectors targeting EHMT1 were 5'-CGC TAT GAT GAT GAT GAA TAA-3' (shEHMT1-1) and 5'-GAG GAT AGT AGG ACT TCT AAA-3' (shEHMT1-2). The corresponding double-stranded DNA sequences were ligated into pSUPER-Retro (GFP-Neo) (Oligoengine) for retroviral expression. For retrovirus production, Phoenix packaging cells were transfected at 70% confluence by calcium phosphate method with 10 μ g retroviral vectors. After 48 h, the viral supernatant was collected and filtered. Cells were incubated overnight with the viral supernatant and supplemented with 6 μ g ml⁻¹ polybrene. Subsequently, puromycin (PRDM16 and EHMT1), hygromycin (C/EBP- β) or G418 (shRNAs) were used for selection.

Gene expression analysis. Total RNA was isolated from tissues using Trizol (Invitrogen) or RiboZol reagents (AMRESCO) following the manufacturers' protocols. Quality of RNA from all the samples was checked by spectrophotometer. Reverse transcription reactions were performed using an iScript complementary DNA (cDNA) synthesis kit (Bio-Rad). The sequences of primers used in this study can be found in Supplementary Table 2. Quantitative reverse transcriptase PCR (qRT–PCR) was performed with SYBR green fluorescent dye using an ABI ViiA7 PCR machine. Relative mRNA expression was determined by the $\Delta\Delta$ -Ct method using TATA-binding protein as an endogenous control to normalize samples.

RNA-sequencing and gene ontology analysis. Total RNA was isolated from the presumptive interscapular BAT depots of WT and *Ehmt1^{myf5}* knockout mice at P1 or from the interscapular BAT depots of WT and *Ehmt1^{adipo}* knockout mice at 12 weeks old. RNA-sequencing libraries were constructed from 50 ng of total RNA from the *Ehmt1^{adipo}* knockout and *Ehmt1^{myf5}* knockout BAT using an Ovation RNA-sequencing system version 2 (NuGEN). mRNA was reverse transcribed to cDNAs using a combination of random hexamer and a poly-T chimeric primer. The cDNA libraries were subsequently amplified by single primer isothermal amplification³⁰ using an Ultralow DR library kit (NuGEN) according to the manufacturer's instructions. The qualities of the libraries were determined by Bioanalyzer (Agilent Technologies). Subsequently, high-throughput sequencing was performed using a HiSeq 2500 instrument (Illumina) at the University of California, San Francisco Genomics Core Facility. RNA-sequencing reads for each library were mapped independently using TopHat version 2.0.8 against the University of California, Santa Cruz (UCSC) mouse genome build mm9 indexes, downloaded from the TopHat website (<http://tophat.cbcb.umd.edu/igenomes.shtml>). The mapped reads were converted to fragments per kilobase of exon per million fragments mapped by running Cuffdiff 2 (ref. 31) on the alignments from TopHat and the UCSC coding genes to estimate amounts of gene and isoform expression. Based on the list of genes that showed significant difference ($P < 0.05$, the delta-method-based hypothesis test) from the RNA-sequencing data, enrichment of the Gene Ontology biological process terms (GO FAT category) was analysed using the Gene Set Enrichment Analysis (GSEA) program, according to the method described by the previous paper³². RNA-sequencing reads have been deposited in ArrayExpress (www.ebi.ac.uk) under accession number E-MTAB-1704.

Immunocytochemistry. Differentiated C2C12 myotubes or COS7 cells expressing green fluorescent protein (GFP)–PRDM16 and EHMT1 constructs were fixed with 4% paraformaldehyde for 10 min at room temperature (24 °C), rinsed with PBS and then exposed to 0.2% Triton X-100 in PBS for 5 min. The cells were subsequently incubated with anti-MF20 mouse antibody (DSHB, 1:50) for MHC and with Flag antibody (M2, 1:200) for EHMT1. After washing with PBS, Alexa 594-labelled anti-mouse IgG (1:800) was added as a secondary antibody.

Protein interaction analyses. Immortalized brown fat cells stably expressing Flag-tagged WT, PR-domain deletion mutant and ZF-1 deletion mutant of PRDM16 or an empty vector were grown to confluence². Nuclear extracts were isolated from these cells and incubated with Flag M2 agarose beads, washed in a binding buffer (180 mM KCl) and subsequently eluted either by 3 \times or by 1 \times Flag peptides (0.2 μ g ml⁻¹). The eluted proteins were subjected to histone methyltransferase assay or to reverse-phase LC–MS/MS for peptide sequencing using a high-resolution hybrid mass spectrometer (LTQ-Orbitrap, Thermo Scientific) with TOP10 method. Data obtained was annotated using the IPI mouse database³³. Proteins were considered significantly identified with at least two unique valid peptides, and the false

discovery rate was estimated to be 0% using the target-decoy approach³⁴. To confirm the interaction between PRDM16 and EHMT1 in brown adipocytes, the immunopurified complex was purified using anti-EHMT1 (R&D Systems) or Flag antibody (M2) and subjected to 4–12% SDS–PAGE. Rabbit polyclonal PRDM16 antibody³⁵ or EHMT1 antibody (R&D Systems) was used for western blotting. COS7 cells expressing haemagglutinin (HA)-tagged PRDM16 or deletion fragments of Flag-tagged EHMT1²⁹ were collected 48 h after transfection. Total cell lysates were incubated overnight at 4 °C with Flag M2 agarose beads, washed and eluted by boiling. The immunoprecipitants were analysed by western blot analysis using HA antibody (Roche). For *in vitro* binding assays, various fragments of the GST–fusion PRDM16 fragments were purified as previously described¹⁶. ³⁵S-labelled proteins (EHMT1, EHMT2, CtBP1, C/EBP- β) were prepared with a TNT reticulocyte lysate kit (Promega). Equal amounts of GST fusion proteins (2 μ g) were incubated overnight at 4 °C with *in vitro* translated proteins in a binding buffer containing 20 mM HEPES pH 7.7, 300 mM KCl, 2.5 mM MgCl₂, 0.05% NP40, 1 mM DTT and 10% glycerol. The sepharose beads were then washed five times with the binding buffer. Bound proteins were separated by SDS–PAGE and analysed by autoradiography.

Histone methylation assay. The PRDM16 transcriptional complex was immunopurified from nuclear extracts of brown adipocytes using Flag M2 agarose or IgG (negative control). The immunoprecipitants were incubated with 2 μ g of core histone (Millipore) with [³H]S-adenosyl-methionine at 30 °C for 1 h. Subsequently, the reaction was stopped by addition of sample buffer. Core histone was resolved by 4–12% SDS–PAGE and detected by autoradiography or by scintillation counter.

Chromatin immunoprecipitation assay. After cross-linking with 1% formaldehyde at room temperature (24 °C) for 10 min, total cell lysates from brown adipocytes were sonicated to shear the chromatin, and immunoprecipitated overnight at 4 °C using antibodies for H3 di-methyl and tri-methyl K9 (Abcam), acetyl-H3K9/K14 (Millipore), pan-H3 (Cell Signaling), EHMT1 (R&D Systems) or IgG (Santa Cruz). After extensive washing, the immunoprecipitants were eluted with 2% SDS in 0.1 M NaH₂CO₃. Cross-linking was reversed by heating at 65 °C overnight. Input DNA and immunoprecipitated DNA were purified by a PCR purification kit (Qiagen) and analysed by qRT–PCR using SYBR green fluorescent dye (Bio-Rad). Enrichment of each protein was calculated as a ratio to input DNA. Primer sequences used in the chromatin immunoprecipitation assays are provided in Supplementary Table 2.

Protein stability assay. COS7 cells expressing HA-tagged PRDM16 and EHMT1 or vector control were incubated with a medium containing 60 μ g ml^{−1} cycloheximide for up to 24 h. Total cell lysates were isolated and separated by SDS–PAGE. Horseradish peroxidase-conjugated HA antibody (Sigma) and β -actin (Sigma) were used for western blotting. Image J software was used to quantify the intensity of signals. Half-life of the protein was estimated by regression analysis.

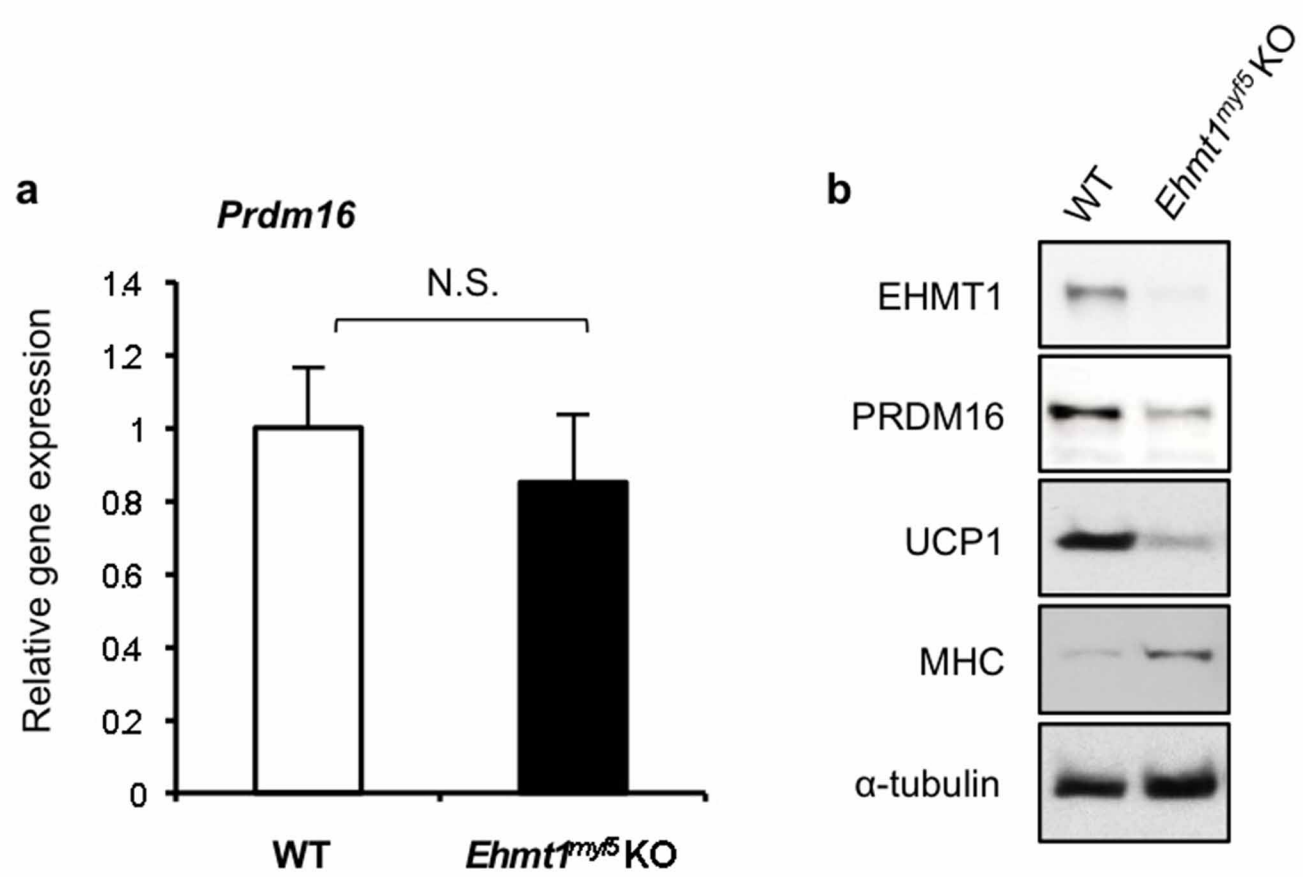
Reporter gene assay. A luciferase reporter gene controlled by PPAR- γ binding sites (3 \times DR1-Luciferase) was transiently transfected with PPAR- γ /RXR- α , PRDM16

and EHMT1 expression plasmids in COS7 cells using Lipofectamine 2000 (Invitrogen). Forty-eight hours after the transfection, cells were collected and reporter gene assays used the Dual Luciferase Kit (Promega). Transfection efficiency was normalized by measuring expression of *Renilla* luciferase.

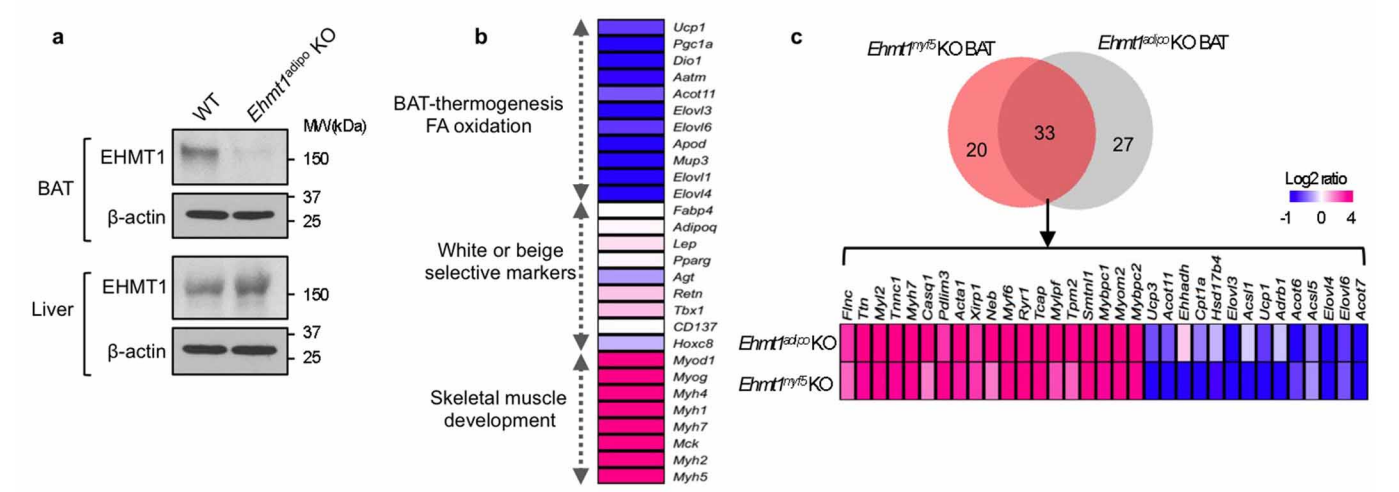
Cellular respiration assay. Immortalized brown adipocytes were transduced with retroviral shEHMT1 (shEHMT1-1) or scramble control and induced to differentiate. Brown adipocytes expressing EHMT1 or vector control were also differentiated under a pro-adipogenic condition. At day 6 of differentiation, oxygen consumption was measured as previously described³⁶. Oligomycin was used to determine uncoupled respiration. In addition, antimycin A was added at the end of experiments to determine non-mitochondrial cellular respiration. For cAMP-induced respiration assays, fully differentiated brown adipocytes were incubated with 0.5 mM dibutyryl cyclic AMP for 12 h before measuring oxygen consumption.

Statistical analyses. Statistical analysis used JMP version 9.0 (SAS Institute). Two-way repeated-measures analysis of variance was applied to determine the statistical difference in glucose tolerance test, insulin tolerance test, body mass gain and rectal temperatures between genotypes. Effect size and power analysis were done by the pwr.t.test function of the R statistics package. Other statistical comparisons were assessed by an unpaired Student's *t*-test. *P* < 0.05 was considered significant throughout the study.

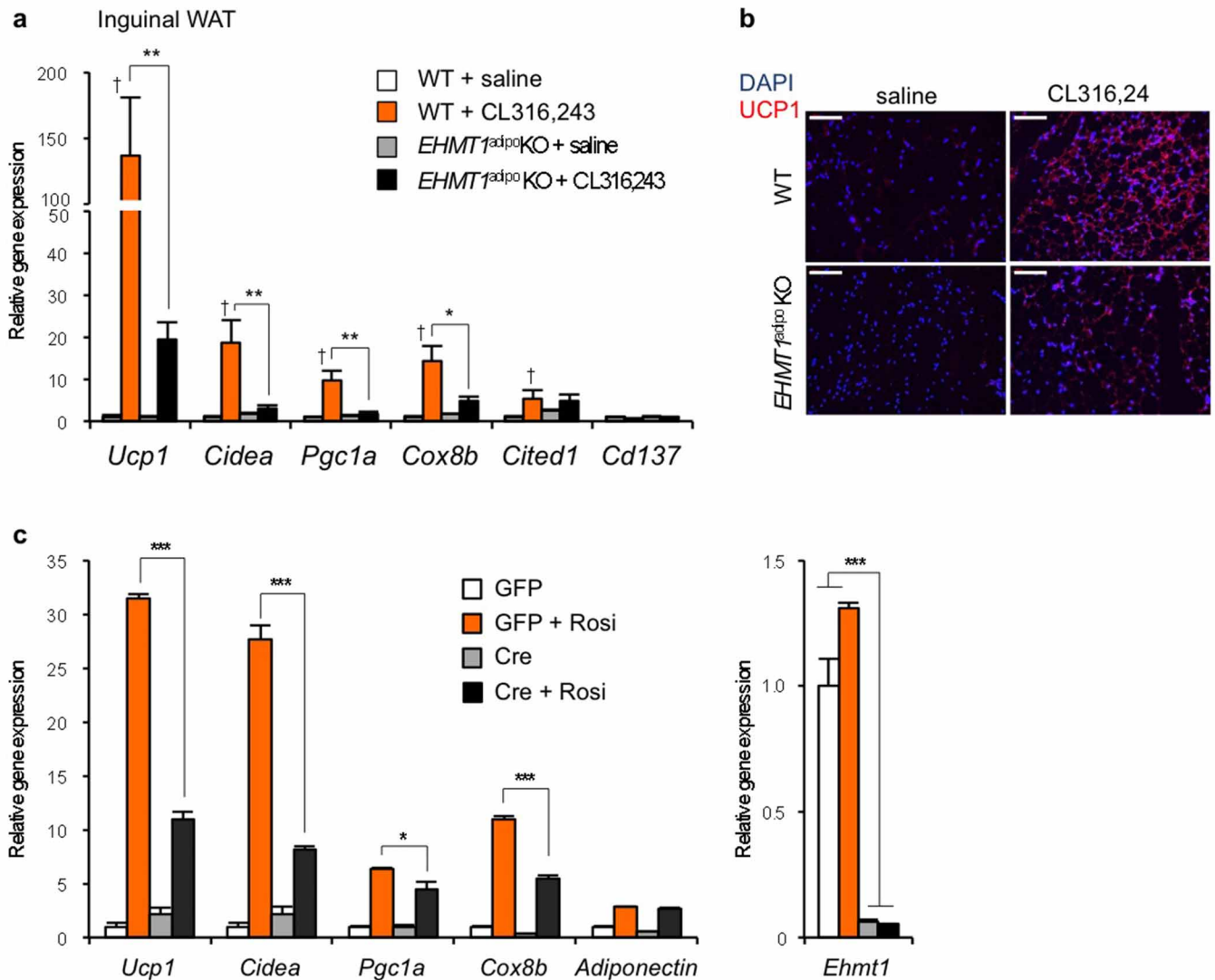
26. Huh, M. S., Parker, M. H., Scime, A., Parks, R. & Rudnicki, M. A. Rb is required for progression through myogenic differentiation but not maintenance of terminal differentiation. *J. Cell Biol.* **166**, 865–876 (2004).
27. Mao, X. *et al.* APPL1 binds to adiponectin receptors and mediates adiponectin signalling and function. *Nature Cell Biol.* **8**, 516–523 (2006).
28. Liisberg Aune, U., Ruiz, L. & Kajimura, S. Isolation and differentiation of stromal vascular cells to beige/brite cells. *J. Visual. Exp.* e50191 (2013).
29. Takahashi, A. *et al.* DNA damage signaling triggers degradation of histone methyltransferases through APC/C(Cdh1) in senescent cells. *Mol. Cell* **45**, 123–131 (2012).
30. Kurn, N. *et al.* Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clin. Chem.* **51**, 1973–1981 (2005).
31. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnol.* **31**, 46–53 (2013).
32. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
33. Kersey, P. J. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988 (2004).
34. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).
35. Seale, P. *et al.* Transcriptional control of brown fat determination by PRDM16. *Cell Metab.* **6**, 38–54 (2007).
36. Ohno, H., Shinoda, K., Spiegelman, B. M. & Kajimura, S. PPAR γ agonists induce a white-to-brown fat conversion through stabilization of PRDM16 protein. *Cell Metab.* **15**, 395–404 (2012).



Extended Data Figure 1 | EHMT1 regulates endogenous PRDM16 protein expression *in vivo*. **a**, The putative BAT was micro-dissected from WT and *Ehmt1^{myf5}* knockout embryos. mRNA expression of *Prdm16* was measured by qRT-PCR. Data are presented as mean and s.e.m. ($n = 8-10$). **b**, Western blotting to detect endogenous EHMT1, PRDM16, UCP1 and MHC in BAT from WT and *Ehmt1^{myf5}* knockout embryos. α -Tubulin protein was shown as a loading control.



Extended Data Figure 2 | Ectopic activation of skeletal-muscle-selective genes and reduction of BAT-selective genes in the BAT from *Ehmt1^{adipo}* knockout mice. **a**, Western blotting for endogenous EHMT1 in BAT and liver from WT and *Ehmt1^{adipo}* knockout mice. β -Actin protein was shown as a loading control. **b**, Amounts of mRNA expression of BAT, skeletal muscle, white fat and beige-fat selective genes in BAT from *Ehmt1^{adipo}* knockout mice. Values were normalized to those in WT mice. The amounts of mRNA were visualized by a heat-map using Multi Experiment Viewer. **c**, Venn diagram showing the overlapped genes between *Ehmt1^{myf5}* knockout and *Ehmt1^{adipo}* knockout mice. RNA-sequencing and gene ontology analyses identified 33 genes that were similarly dysregulated both in the *Ehmt1^{myf5}* knockout BAT and the *Ehmt1^{adipo}* knockout BAT. The mRNA expression values were normalized to WT mice for each knockout model and visualized by a heat-map using Multi Experiment Viewer. The colour scale shows the amounts of mRNA of the genes in a blue (low)-white (no change)-red (high) scheme.



Extended Data Figure 3 | EHMT1 is required for beige/brite cell development. **a**, The β_3 -AR agonist CL316,243 at a dose of 0.5 mg kg^{-1} or saline were administered to WT or *Ehmt1^{adipo}* knockout mice for 7 days. Inguinal WAT was collected for gene expression analysis. Amounts of mRNA expression of BAT and beige-fat selective genes (as indicated) were measured by qRT-PCR ($n = 3$ –6). †Significant between saline and CL316,243 in WT mice. **b**, Immunohistochemistry for UCP1 in **a**. Scale bar, $100 \mu\text{m}$. Nuclei were stained with DAPI. **c**, To test a cell-autonomous requirement for EHMT1 in

beige/brite cell development, the stromal vascular (SV) fractions were isolated from the inguinal WAT of *Ehmt1^{flox/flox}* mice. Cells were infected with adenovirus expressing GFP or Cre. The SV cells were differentiated in the presence or absence of rosiglitazone (Rosi) at $0.5 \mu\text{M}$. Amounts of mRNA expression of BAT-selective genes (as indicated) were measured by qRT-PCR. Deletion of *Ehmt1* was confirmed by qRT-PCR (right graph) ($n = 3$); data are presented as mean and s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Structural basis for the modular recognition of single-stranded RNA by PPR proteins

Ping Yin^{1,2*}, Quanxiu Li^{1,2*}, Chuangye Yan^{2,3}, Ying Liu^{1,2}, Junjie Liu^{2,3}, Feng Yu⁴, Zheng Wang⁵, Jiafu Long⁵, Jianhua He⁴, Hong-Wei Wang^{2,3}, Jiawei Wang^{1,2}, Jian-Kang Zhu^{6,7}, Yigong Shi^{2,3} & Nieng Yan^{1,2}

Pentatricopeptide repeat (PPR) proteins represent a large family of sequence-specific RNA-binding proteins that are involved in multiple aspects of RNA metabolism. PPR proteins, which are found in exceptionally large numbers in the mitochondria and chloroplasts of terrestrial plants^{1–5}, recognize single-stranded RNA (ssRNA) in a modular fashion^{6–8}. The maize chloroplast protein PPR10 binds to two similar RNA sequences from the *ATPI-ATPH* and *PSAJ-RPL33* intergenic regions, referred to as *ATPH* and *PSAJ*, respectively^{9,10}. By protecting the target RNA elements from 5' or 3' exonucleases, PPR10 defines the corresponding 5' and 3' messenger RNA termini^{9–11}. Despite rigorous functional characterizations, the structural basis of sequence-specific ssRNA recognition by PPR proteins remains to be elucidated. Here we report the crystal structures of PPR10 in RNA-free and RNA-bound states at resolutions of 2.85 and 2.45 Å, respectively. In the absence of RNA binding, the nineteen repeats of PPR10 are assembled into a right-handed superhelical spiral. PPR10 forms an antiparallel, intertwined homodimer and exhibits considerable conformational changes upon binding to its target ssRNA, an 18-nucleotide *PSAJ* element. Six nucleotides of *PSAJ* are specifically recognized by six corresponding PPR10 repeats following the predicted code. The molecular basis for the specific and modular recognition of RNA bases A, G and U is revealed. The structural elucidation of RNA recognition by PPR proteins provides an important framework for potential biotechnological applications of PPR proteins in RNA-related research areas.

PPR proteins function in multiple aspects of organelle RNA metabolism, such as RNA splicing, editing, degradation and translation^{1–5}. In plants, PPR mutants may cause embryonic lethality^{12–14}, and a number of PPR proteins act as restorers of fertility to overcome cytoplasmic male sterility^{15–19}. In humans, mutations in the mitochondrial PPR protein LRPPRC are associated with the French-Canadian-type Leigh syndrome characterized by the deficiency in Complex IV^{20,21}.

PPR proteins contain 2–30 tandem repeats, each typically comprising 35 amino acids that are organized into a hairpin of α -helices^{1,6,22,23}. PPRs are divided into two classes: the P-class, whose members only comprise the 35-amino-acid repeats; and the PLS-class, which has repeats of 31–36 amino acids and extra domains at the carboxyl terminus^{3,12}. Computational and biochemical analyses suggest that PPR proteins may recognize RNA in a modular fashion, but different from that of the RNA-binding PUF domain^{6,24}. The putative RNA recognition code by PPR proteins derived from bioinformatic and biochemical analyses awaits structural corroboration^{2,6–8}.

To elucidate the mechanism of specific RNA recognition by PPR proteins, we sought to determine the crystal structure of well-characterized PPR proteins in complex with their target RNAs. The recombinant protein of maize chloroplast PPR10, which belongs to the P-class, specifically binds to the 17-nucleotide (nt) (*ATPH*) and 18-nt (*PSAJ*) RNA

oligonucleotides (Extended Data Fig. 1a)¹⁰. We launched a systematic effort to determine the structures of PPR10 in both RNA-free and RNA-bound states.

The crystal structure of the RNA-free PPR10 fragment (residues 61–786) containing quadruple Cys mutations (C256S/C279S/C430S/C449S) was determined at 2.85 Å resolution. PPR10 forms a right-handed two-turn superhelical assembly, with 19 PPR motifs (residues 107–771) capped by three short α -helices at the amino-terminal domain (NTD) and a single α -helix at the C terminus (Fig. 1a). Capping motifs are known to contribute to ligand specificity for repeat proteins such as TPR (tetra-tricopeptide repeat)²⁵ and TALE (transcription activator-like effector)^{26,27}. The function of the extra motifs in PPR10 remains to be determined.

The 35 amino acids in each PPR motif form a hairpin of α -helices, each containing four helical turns, followed by a five-residue loop (Fig. 1b). The two helices, designated helix a and helix b, are connected by a short turn of two amino acids. Helices a and b of each repeat constitute the inner and outer layers of the superhelical assembly, respectively (Fig. 1a). In the crystals, there is one molecule of PPR10 in each asymmetric unit, yet two symmetry-related molecules are intertwined in an antiparallel fashion. The N terminus of one molecule is in close contact with the C terminus of the other, yielding an overall appearance of an ellipsoid with a polar axis of approximately 140 Å and an equatorial diameter of 70 Å (Fig. 1c).

On the basis of the PPR10 structure, we defined the starting amino acid of helix a as the first residue in a PPR motif (Fig. 1b and Extended Data Fig. 1b). This definition results in a one-residue shift either forwards^{6,12} or backwards^{7,28} within each repeat compared to the previously described boundary of a PPR motif (Extended Data Fig. 1c). With the new boundary assignment of a PPR motif, the residues that were predicted to determine RNA binding specificity are all included in one structurally intact motif. We hope that this structure-based demarcation of the PPR motif will simplify future descriptions of PPR proteins.

After numerous unsuccessful crystallization trials for PPR10–*ATPH* complexes, we finally determined the structure of PPR10 (residues 69–786, C256S/C279S/C430S/C449S) in the presence of 18-nt *PSAJ* RNA (5'-GUAUUCUUUAUUAUUUC-3') at 2.45 Å resolution (Extended Data Table 1). In the crystals, there is one antiparallel PPR10 dimer in each asymmetric unit. Analysis by sedimentation equilibrium analytical ultracentrifugation (SE-AUC) of *PSAJ*-bound PPR10 (residues 37–786, C256S/C279S/C430S/C449S) supports its dimeric existence at micromolar concentration in solution (Extended Data Fig. 2). The two PPR10 protomers can be superimposed with a root-mean-squared deviation of 1.3 Å over 629 C α atoms (Extended Data Fig. 3). The overall appearance of the dimer has changed to a hollow cylindrical tube (Fig. 2a), and the N- and C-terminal portions of the PPR10 protomer are compressed towards the centre, resulting in a reduction of 20 Å in axial length (Fig. 2b).

¹State Key Laboratory of Bio-membrane and Membrane Biotechnology, Tsinghua University, Beijing 100084, China. ²Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China. ³Ministry of Education Key Laboratory of Protein Science, Tsinghua University, Beijing 100084, China. ⁴Shanghai Institute of Applied Physics, Chinese Academy of Sciences, 239 Zhangheng Road, Shanghai 201204, China. ⁵State Key Laboratory of Medicinal Chemical Biology and College of Life Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China. ⁶Shanghai Center for Plant Stress Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China. ⁷Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, Indiana 47907, USA.

*These authors contributed equally to this work.

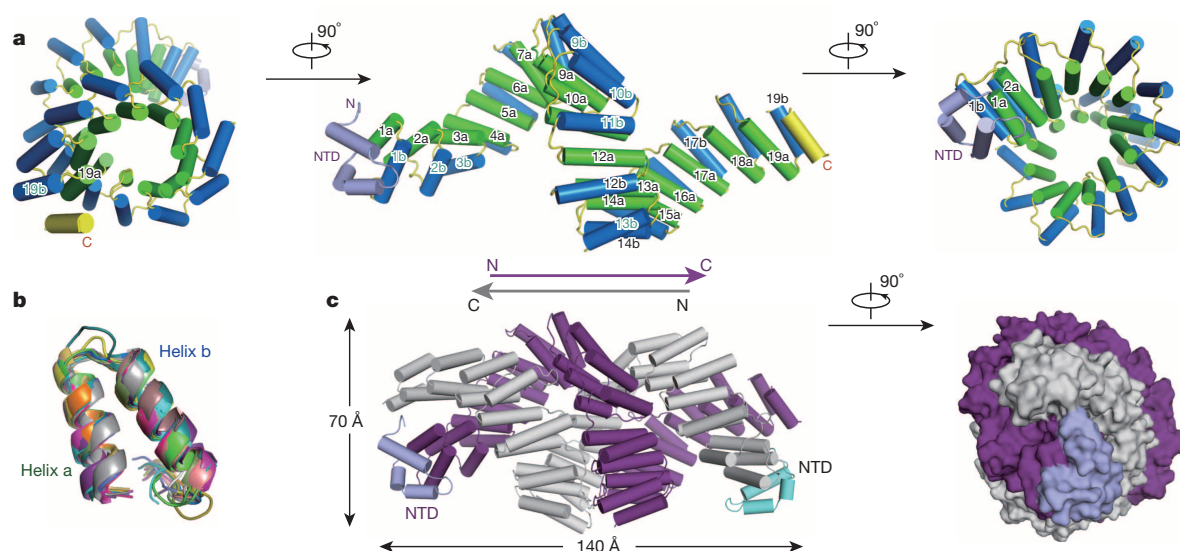


Figure 1 | Crystal structure of RNA-free PPR10. **a**, Overall structure of RNA-free PPR10. The fragment (residues 61–786, C256S/C279S/C430S/C449S) comprises 19 repeats capped by a small NTD (light purple) and a C-terminal helix (yellow). The two helices within each repeat, designated helix

a and helix b, are coloured green and blue, respectively. **b**, Structural superimposition of the 19 repeats of PPR10. **c**, Overall structure of the PPR10 dimer. Two molecules from adjacent asymmetric units form an intertwined antiparallel dimer. All structure figures were prepared with PyMol³⁰.

Following assignment of most amino acids of PPR10 into the electron density map, strong electron densities indicative of RNA bases became clearly visible in the cavities on both ends of the cylindrical tube (Fig. 2c). Assignment of 18 and 14 nucleotides of the two bound RNA elements was validated by the anomalous signals of bromine (Br), which were collected for crystals of PPR10 bound to Br-labelled RNA oligonucleotides (Extended Data Fig. 4 and Extended Data Table 2). The 5' and 3' portions of the ssRNA are specifically recognized by the N-terminal repeats of one protomer and C-terminal repeats of the other. By contrast, the middle portion of the ssRNA, comprising nucleotides U5 to A10, remains largely uncoordinated by PPR10 (Fig. 2d and Extended Data Fig. 5a, b).

PPR10 has 19 repeats and the bound *PSAJ* RNA contains 18 nucleotides. Consistent with a bioinformatic prediction⁶, specific recognition of the *PSAJ* RNA begins with repeat 3 (Fig. 3a). Each of the first four nucleotides on the 5' end, 5'-GUAU-3', is recognized by one PPR10 repeat. Such recognition exhibits a modular pattern involving residues that were predicted through biochemical and bioinformatic analyses^{2,6,7}. Each RNA base is surrounded by four residues, the 2nd residues from two adjacent repeats, and the 5th and 35th residues from a corresponding repeat. In addition to base recognition, the backbone phosphate or ribose groups of the bound *PSAJ* RNA are also coordinated by charged or polar amino acids from PPR10 (Extended Data Fig. 5c).

A polar amino acid located at the 5th position in each repeat appears to be the most important determinant for RNA base specificity. Thr 178, Asn 213, Ser 249 and Asn 284 in repeats 3–6 recognize the bases G1, U2, A3 and U4, respectively, through direct hydrogen bonds (Fig. 3b). The importance of the 5th residue in RNA recognition is supported by mutational analysis. Mutating any of the 5th residues in repeats 4 (N213A), 5 (S249L) or 6 (N284A) resulted in complete abolishment of RNA binding. By contrast, substitution of the 5th residues of repeats 7, 8, 10, 11 or 13, which are not involved in RNA binding in the structure, showed little or no effect on *PSAJ* binding (Extended Data Fig. 6).

Buttressing the hydrogen bonds, five residues at the 2nd position of PPR repeats 3–7 sandwich the four bases mainly through van der Waals interactions (Fig. 3a). For example, G1 is surrounded by Arg 175/Val 210 of repeats 3 and 4. Similarly, U2, A3 and U4 are sandwiched by Val 210/Phe 246, Phe 246/Val 281 and Val 281/Val 316, respectively (Fig. 3). The 35th residue is located in the vicinity of the base. It is possible that water molecules, although invisible in the structure, may mediate hydrogen bonds between the polar residues and the bases. Importantly, Asp 244

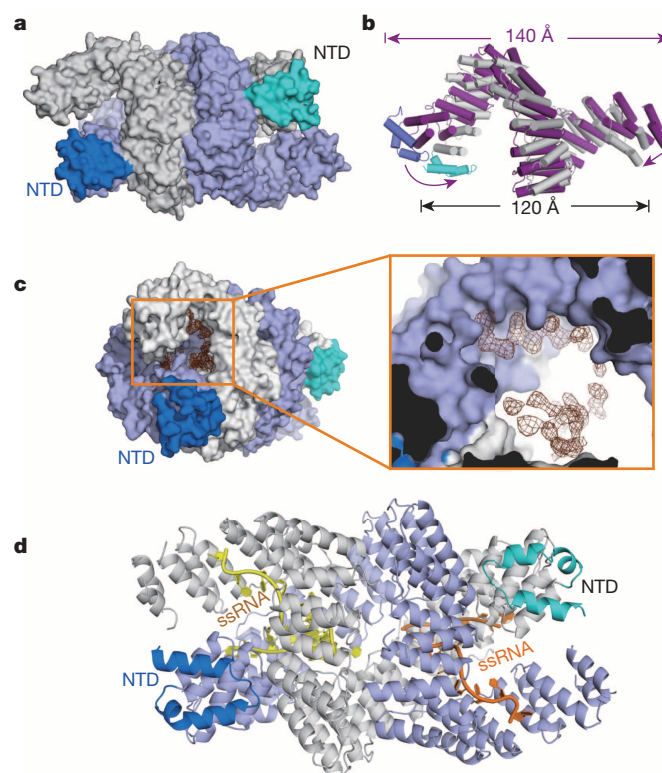


Figure 2 | Structure of PPR10 bound to an 18-nt *PSAJ* RNA element. **a**, The PPR10 dimer (residues 69–786, C256S/C279S/C430S/C449S) forms a cylindrical tube in the presence of *PSAJ*. The two protomers are coloured light purple and grey with their NTDs coloured blue and cyan. **b**, The PPR10 protomer undergoes pronounced conformational changes upon binding to *PSAJ*. The structure of RNA-free PPR10 is coloured magenta with the NTD coloured lilac. **c**, Electron densities found in the cavities on both ends of the PPR10 dimer. The 'omit' electron density, with a close-up view in the inset, is contoured at 3σ . **d**, Overall structure of the PPR10–*PSAJ* complex. The two ssRNA molecules are coloured yellow and orange.

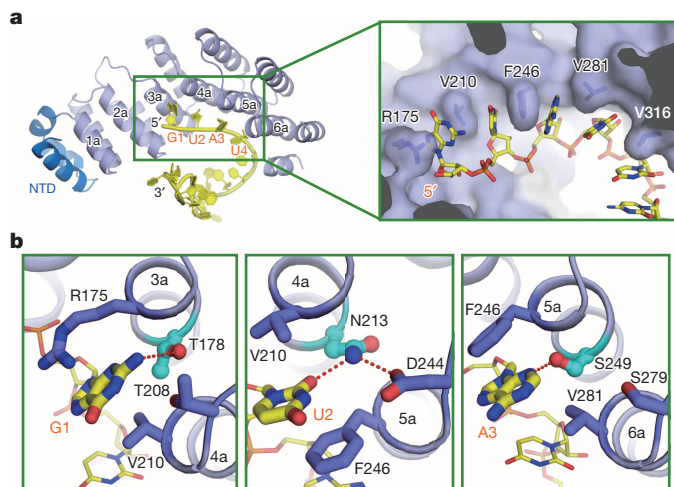


Figure 3 | Base-specific recognition of ssRNA by PPR10 repeats. **a**, The four nucleotides at the 5' end of the PSAJ RNA segment are specifically recognized in a modular fashion. Inset: each of the four RNA bases at the 5' end is sandwiched by two residues at the 2nd positions of adjacent repeats. **b**, Specific recognition of the bases G, U and A by PPR10 repeats. The side chain of the 5th residue in each repeat, which makes a direct hydrogen bond to the base, is highlighted in cyan. The hydrogen bonds are represented by red dotted lines.

and Asp 314, the 35th residues in repeats 4 and 6, are respectively hydrogen bonded to Asn 213 and Asn 284, the 5th residues in the corresponding repeats, and may help to stabilize their conformation for base recognition (Fig. 3b and Extended Data Fig. 6d).

Recognition of the 3' end of the PSAJ RNA by the C-terminal repeats in the other PPR10 protomer appears to be less modular except for U15 and U16, which are coordinated by repeats 16 and 17 following the described recognition pattern for U2 and U4 (Fig. 4a). The bases A11 and C18 are coordinated in a non-modular fashion. The adenine base of A11 donates a hydrogen bond to Asp 630, the 35th residue of repeat 15, whereas the cytosine base of C18 makes a hydrogen bond to Ser 714 on the last helical turn of helix 18a (Fig. 4b). A number of direct and water-mediated hydrogen bonds are found between the backbone phosphate/ribose groups of U15–U16–U17–C18 and the polar residues on repeats 15–19 of PPR10 (Fig. 4c).

In the structure of PSAJ-bound PPR10, only six out of 18 nucleotides in the PSAJ RNA element strictly follow the modular pattern (Fig. 4d). Two bases, A11 and C18, are bound by PPR10 in a non-modular fashion, and the other 10 bases are literally uncoordinated. Notably, only 17 repeats in each PPR10 protomer are available for the binding of 18 nucleotides. It remains to be seen whether the 17-nt *ATPH* binds to repeats 3–19 of PPR10 in a completely modular fashion. Binding of *ATPH* leads to dissociation of the PPR10 dimer (Extended Data Fig. 2)^{6,10}. Interestingly, analysis by SE-AUC suggests that, although RNA-free PPR10 forms a stable dimer, PSAJ binding weakens the PPR10 dimer formation (Extended Data Fig. 2). It remains to be investigated whether PSAJ-bound PPR10 is a dimer under physiological conditions, in which the protein concentration can be very low. Nevertheless, the crystal structure of PPR10 bound to the 18-nt PSAJ RNA element reveals the molecular basis for specific recognition between a PPR protein and its target RNA sequence.

Recognition of the six RNA nucleotides 5'-G1-U2-A3-U4-3' and 5'-U15-U16-3' by repeats 3–6 and repeats 16 and 17, respectively, largely supports the predicted code for base discrimination in ssRNA, where the bases G, U and A are specifically recognized by the 5th residues of a PPR motif: Thr, Asn and Ser, respectively^{6,7}. The 2nd and 35th residues, sitting in the vicinity of the bases, contribute to RNA binding (Fig. 4d and Extended Data Fig. 6c, d). The prediction that base C is recognized by an Asn at the 5th position can also be conveniently rationalized based on our crystal structure (Extended Data Fig. 7). One unexpected

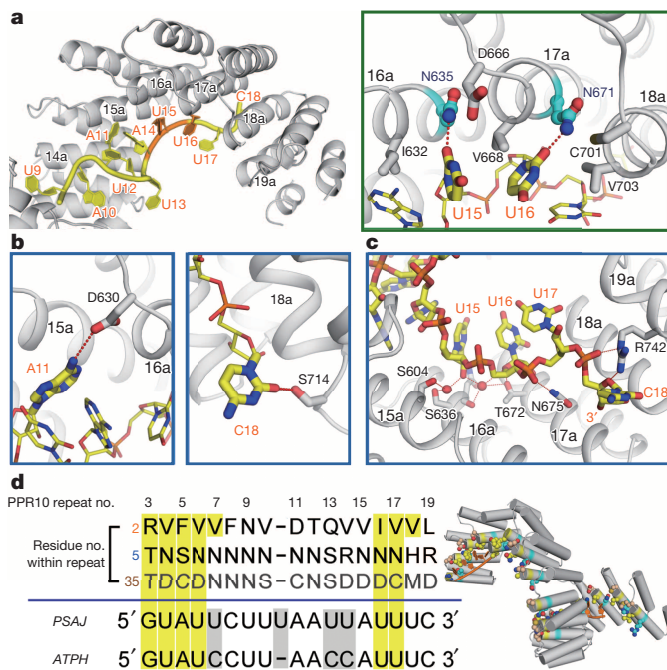


Figure 4 | Coordination of the 3'-end segment of the PSAJ RNA by PPR10. **a**, Recognition of the bases U15 and U16 by PPR10 follows the code discussed in Fig. 3. **b**, The bases A11 and C18 are hydrogen bonded to polar residues on PPR10 in a non-modular fashion. **c**, The backbone of the 3'-end segment of PSAJ is coordinated through direct or water-mediated hydrogen bonds. **d**, Summary of specific recognition of PSAJ RNA by PPR10. Left, the residues at the 2nd, 5th and 35th positions of PPR10 motifs and the corresponding sequences of the target RNA elements. The structurally corroborated recognition codes are shaded yellow. The difference in RNA sequences of PSAJ and ATPH is shaded grey. Right, the structure of repeats 3–19 of PPR10 with their 2nd (yellow), 5th (cyan) and 35th (wheat) residues shown in spheres, and the six recognized RNA bases shown in orange.

feature is that the 2nd residues from two consecutive repeats sandwich one base; therefore the identity of the 2nd residue on the next repeat must be considered for RNA binding. Base sandwiching by hydrophobic residues or Arg is also observed in the recognition of ssRNA by PUF proteins^{24,29}, although PPR and PUF proteins exhibit distinct RNA binding modes (Extended Data Fig. 8).

Further biochemical, computational, structural and *in vivo* characterizations are required to completely rationalize the codes for specific RNA recognition by PPRs and to engineer PPR proteins for targeted RNA manipulations. The structures reported here provide unprecedented insights into the recognition mechanism of RNA elements by PPR proteins and serve as an important foundation for understanding the function and mechanism of numerous PPR proteins in RNA metabolism, and for the potentially customized design of specific-RNA-binding PPR proteins.

METHODS SUMMARY

The codon-optimized complementary DNA of full-length PPR10 (Gene ID: 100302579) from *Zea mays* was subcloned into pET15b vector (Novagen). Overexpression of PPR10 protein was induced in *Escherichia coli* BL21(DE3). To crystallize PPR10, we mounted a systematic protein engineering effort including a series of protein truncations and mutations of Cys residues. There are 18 Cys residues within the repeat region. We generated 18 mutants, each consisting of a single Cys to Ser mutation and tested their binding with the 17-nt *ATPH* element. For those that completely retained binding affinity, we further grouped them to double, triple and quadruple mutations. Finally, the PPR10 mutant containing C256S/C279S/C430S/C449S showed the same binding affinity as wild type and exhibited excellent protein behaviour. All the PPR10 proteins used in the manuscript contain the quadruple Cys mutations. The RNA-free PPR10 fragment (residues 61–786, C256S/C279S/C430S/C449S) was eventually crystallized in the space group P2₁2₁2. The structure

was determined by selenium-based single-wavelength anomalous diffraction and refined to 2.85 Å resolution (Extended Data Table 1). In the effort to crystallize PPR10 in complex with its target RNA, despite numerous trials, most PPR10-*ATPH* complexes defied crystallization; for those that crystallized, X-ray diffraction was consistently poor. We applied the same strategy to complexes between PPR10 and *PSAJ*. After screening more than 100,000 conditions, we were able to crystallize the complex between PPR10 (residues 69–786, C256S/C279S/C430S/C449S) and the 18-nt *PSAJ* RNA (5'-GUAUUCUUUAAUUAUUUC-3') in the space group *P*₄₃. These crystals diffract X-rays beyond 2.5 Å. The structure was determined by molecular replacement using successive segments of the RNA-free PPR10 structure, but not the entire molecule. We were able to assign all 18 nucleotides of one bound *PSAJ* RNA element, but only 14 of the other. For details of electrophoretic mobility shift assay and SE-AUC experiments, please refer to Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 May; accepted 12 September 2013.

Published online 27 October 2013.

- Small, I. D. & Peeters, N. The PPR motif — a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **25**, 45–47 (2000).
- Nakamura, T., Yagi, Y. & Kobayashi, K. Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. *Plant Cell Physiol.* **53**, 1171–1179 (2012).
- Schmitz-Linneweber, C. & Small, I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends Plant Sci.* **13**, 663–670 (2008).
- Fujii, S. & Small, I. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytol.* **191**, 37–47 (2011).
- Kotera, E., Tasaka, M. & Shikanai, T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* **433**, 326–330 (2005).
- Barkan, A. *et al.* A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet.* **8**, e1002910 (2012).
- Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T. & Nakamura, T. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* **8**, e57286 (2013).
- Yagi, Y. *et al.* Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA Biol.* **10**, 1236–1242 (2013).
- Pfalz, J., Bayraktar, O. A., Prikryl, J. & Barkan, A. Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *EMBO J.* **28**, 2042–2052 (2009).
- Prikryl, J., Rojas, M., Schuster, G. & Barkan, A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proc. Natl Acad. Sci. USA* **108**, 415–420 (2011).
- Zhelyazkova, P. *et al.* Protein-mediated protection as the predominant mechanism for defining processed mRNA termini in land plant chloroplasts. *Nucleic Acids Res.* **40**, 3092–3105 (2012).
- Lurin, C. *et al.* Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**, 2089–2103 (2004).
- Cushing, D. A., Forsthoefel, N. R., Gestaut, D. R. & Vernon, D. M. *Arabidopsis emb175* and other *ppr* knockout mutants reveal essential roles for pentatricopeptide repeat (PPR) proteins in plant embryogenesis. *Planta* **221**, 424–436 (2005).
- Khrouchtchova, A., Monde, R. A. & Barkan, A. A short PPR protein required for the splicing of specific group II introns in angiosperm chloroplasts. *RNA* **18**, 1197–1209 (2012).
- Bentolila, S., Alfonso, A. A. & Hanson, M. R. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc. Natl Acad. Sci. USA* **99**, 10887–10892 (2002).
- Desloire, S. *et al.* Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. *EMBO Rep.* **4**, 588–594 (2003).
- Wang, Z. *et al.* Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. *Plant Cell* **18**, 676–687 (2006).
- Chase, C. D. Cytoplasmic male sterility: a window to the world of plant mitochondrial–nuclear interactions. *Trends Genet.* **23**, 81–90 (2007).
- Hu, J. *et al.* The rice pentatricopeptide repeat protein RF5 restores fertility in Hong-Lian cytoplasmic male-sterile lines via a complex with the glycine-rich protein GRP162. *Plant Cell* **24**, 109–122 (2012).
- Mootha, V. K. *et al.* Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA* **100**, 605–610 (2003).
- Ruzzenente, B. *et al.* LRPPRC is necessary for polyadenylation and coordination of translation of mitochondrial mRNAs. *EMBO J.* **31**, 443–456 (2012).
- Howard, M. J., Lim, W. H., Fierke, C. A. & Koutmos, M. Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proc. Natl Acad. Sci. USA* **109**, 16149–16154 (2012).
- Ringel, R. *et al.* Structure of human mitochondrial RNA polymerase. *Nature* **478**, 269–273 (2011).
- Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501–512 (2002).
- Grove, T. Z., Cortajarena, A. L. & Regan, L. Ligand binding by repeat proteins: natural and designed. *Curr. Opin. Struct. Biol.* **18**, 507–515 (2008).
- Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720–723 (2012).
- Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22**, 1716–1720 (2012).
- Kobayashi, K. *et al.* Identification and characterization of the RNA binding surface of the pentatricopeptide repeat protein. *Nucleic Acids Res.* **40**, 2712–2723 (2012).
- Filipovska, A. & Rackham, O. Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol. Biosyst.* **8**, 699–708 (2012).
- DeLano, W. L. The PyMOL Molecular Graphics System. <http://www.pymol.org> (2002).

Acknowledgements We thank X. Yu and Y. Chen at the Institute of Biophysics, Chinese Academy of Sciences, for technical support. We thank K. Hasegawa and T. Kumasaka at the SPring-8 beamline BL41XU for on-site assistance. This work was supported by funds from the Ministry of Science and Technology (grant number 2011CB910501 for N.Y.), and Projects 91017011 (N.Y.), 31070644 (N.Y.), 31021002 (Y.S., N.Y., J.W.) and 31200567 (P.Y.) of the National Natural Science Foundation of China. The research of N.Y. was supported in part by an International Early Career Scientist grant from the Howard Hughes Medical Institute.

Author Contributions P.Y., Q.L., J.-K.Z., Y.S. and N.Y. designed all experiments. P.Y., Q.L., C.Y., Y.L., J.L., F.Y., Z.W., J.L., J.H., H.-W.W., J.W. and N.Y. performed the experiments. All authors analysed the data and contributed to manuscript preparation. N.Y. wrote the manuscript.

Author Information The atomic coordinates and structure factors of RNA-free and RNA-bound PPR10 have been deposited in the Protein Data Bank (PDB) with the accession codes 4M57 and 4M59, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.Y. (nyan@tsinghua.edu.cn).

METHODS

Protein preparation. The codon-optimized complementary DNA of full-length PPR10 (Gene ID: 100302579) from *Zea mays* was subcloned into pET15b vector (Novagen). Overexpression of PPR10 protein was induced in *E. coli* BL21(DE3) with 0.2 mM isopropyl- β -D-thiogalactoside at an $OD_{600\text{nm}}$ of 1.2. After growing for 16 h at 16 °C, the cells were collected, homogenized in a buffer containing 25 mM Tris-HCl, pH 8.0, and 150 mM NaCl. After sonication and centrifugation, the supernatant was applied to Ni^{2+} affinity resin (Ni-NTA, Qiagen) and further fractionated by ion-exchange chromatography (Source 15Q, GE Healthcare). The PPR10 mutants were generated using two-step PCR and subcloned, overexpressed and purified in the same way as the wild-type protein.

A systematic protein engineering effort was mounted for crystallization of RNA-free and -bound PPR10. A series of protein truncations were tested without giving rise to crystals. There are 18 Cys residues within the repeat region. It is well known that the presence of surface Cys residues, which are subject to oxidation, may lead to protein heterogeneity and impede crystallization. We therefore generated 18 mutants, each consisting of a single Cys to Ser mutation and tested their binding with the 17-nt *ATPH* element. For those that completely retained binding affinity, we further grouped them to double, triple and quadruple mutations. Finally, the PPR10 mutant containing C256S/C279S/C430S/C449S showed the same binding affinity as wild type and exhibited excellent protein behaviour. For consistency, all the PPR10 proteins used in the manuscript contain the quadruple Cys mutations.

For the crystallization trials of RNA-free PPR10 (residues 61–786, C256S/C279S/C430S/C449S), the protein was concentrated and applied to gel filtration chromatography (Superdex-200 10/30, GE Healthcare) in the buffer containing 25 mM Tris-HCl, pH 8.0, 150 mM NaCl and 10 mM dithiothreitol (DTT). Selenomethionine (Se-Met)-derived protein was purified similarly.

To obtain the crystals of protein–RNA complex, PPR10 (residues 69–786, C256S/C279S/C430S/C449S) was purified through Ni^{2+} affinity resin (Ni-NTA, Qiagen), followed by heparin affinity column (HiPrep Heparin FF 16/10, GE Healthcare). The protein was then applied to gel filtration chromatography (Superdex-200 10/30, GE Healthcare). The buffer for gel filtration contained 25 mM Tris-HCl, pH 8.0, 50 mM NaCl, 5 mM MgCl_2 and 10 mM DTT. The peak fractions were incubated with target RNA oligonucleotides with a molar ratio of approximately 1:1.5 at 4 °C for about 40 min before crystallization trials.

Crystallization. Both RNA-free and RNA-bound PPR10 proteins were crystallized by hanging-drop vapour-diffusion method at 18 °C. PPR10 (residues 61–786, C256S/C279S/C430S/C449S), at a concentration of approximately 6.0 mg ml^{-1} , was mixed with an equal volume of reservoir solution containing 1.8–2.1 M sodium formate, and 0.1 M Bis-Tris propane, pH 6.5. Plate-shaped crystals appeared overnight and grew to full size within 1–2 weeks. Se-Met-labelled protein was crystallized similarly.

To obtain crystals of protein–RNA complex, various combinations of protein boundaries and RNA oligonucleotides (Takara) were examined. Because the first visible residue in the structure of RNA-free PPR10 starts at position 69, we invested more effort into this construct. Finally, the protein (residues 69–786, C256S/C279S/C430S/C449S) and 18-nt RNA from the *PSAJ*–*RPL33* intergenic region with the sequence 5′-GUAUUCUUUAAUUAUUUC-3′ (designated *PSAJ* RNA) gave rise to crystals in the reservoir solution containing 8–10% (w/v) polyethylene glycol 3350, 8% Tacsimate, pH 6.0 (Hampton Research), and 0.1 M MES, pH 5.5.

Data collection and structural determination. All data sets were collected at SSRF beamline BL17U or SPring-8 beamline BL41XU and processed with the HKL2000 packages³¹. Further processing was carried out with programs from the CCP4 suite³². Data collection and structure refinement statistics are summarized in Extended Data Tables 1 and 2.

The RNA-free PPR10 structure was solved by single anomalous diffraction (SAD) of Se-Met using the program ShelxC/D/E³³. Then a crude helical model

was manually built in the program Coot³⁴. Using this partial model as input, the identified Se atom positions were refined and phases were recalculated using the SAD experimental phasing module of the program Phaser³⁵. With the improved map, the molecular boundary was unambiguously defined and one molecule was found in an asymmetry unit. The crude model was further rebuilt with Coot and refined with Phenix³⁶. The sequence docking was aided by anomalous map of selenium.

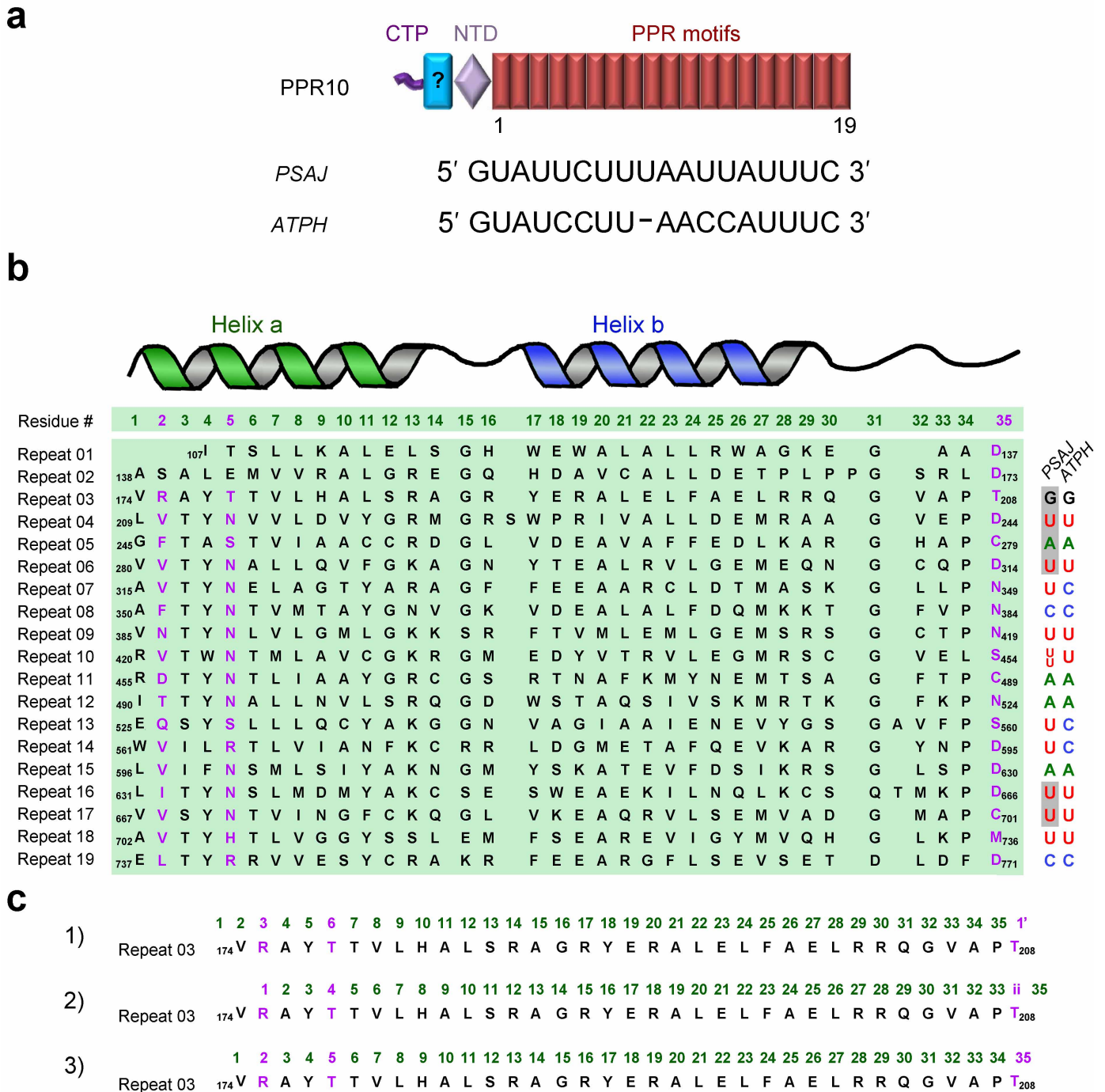
Data sets collected from five crystals of the PPR10–RNA complex were merged for complete and better data. The structure of the PPR10–RNA complex was solved by molecular replacement with the newly solved RNA-free structure as the search model using the program Phaser³⁵. To find the right solution, the structure of the RNA-free PPR10 protomer was divided into three consecutive segments. The assignment of RNA sequence was aided by the anomalous signal of bromine obtained for crystals of PPR10 in complex with Br-labelled RNA oligonucleotides, where U4/U7/U15, U5/U7/U15 or U12 were substituted by 5-bromouracil (Extended Data Table 2). The structure was manually refined with Coot and Phenix iteratively (Extended Data Table 1).

Electrophoretic mobility shift assay (EMSA). The ssRNA oligonucleotides were radiolabelled at the 5′ end with [γ -³²P] ATP (PerkinElmer) catalysed by T4 polynucleotide kinase (Takara). The sequences of ssRNA oligonucleotides used in EMSA are: *PSAJ*, 5′-GUAUUCUUUAAUUAUUUC-3′; and *ATPH*, 5′-GUAUCCUUAACCAUUUC-3′.

For EMSA, PPR10 (residues 37–786, C256S/C279S/C430S/C449S) and the other variants consisting of the indicated point mutations were incubated with approximately 40 pM ³²P-labelled probe in the final binding reactions containing 40 mM Tris-HCl, pH 7.5, 100 mM NaCl, 4 mM DTT, 0.1 mg ml^{−1} BSA, 5 $\mu\text{g ml}^{-1}$ heparin and 10% glycerol at room temperature (22 °C) for 20 min. Reactions were then resolved on 6% native acrylamide gels (37.5:1 for acrylamide:bisacrylamide) in 0.5× Tris-glycine buffer under an electric field of 15 V cm^{−1} for 40 min. Vacuum-dried gels were visualized on a phosphor screen (Amersham Biosciences) with a Typhoon Trio Imager (Amersham Biosciences).

SE-AUC. The oligomeric states of PPR10 (residues 37–786, C256S/C279S/C430S/C449S) with or without target RNA oligonucleotides in solution were investigated by AUC experiments. SE-AUC experiments were performed in a Beckman Coulter XL-I analytical ultracentrifuge using six-channel centrepieces. RNA-free PPR10, *PSAJ*-bound PPR10 and *ATPH*-bound PPR10 were in solutions containing 25 mM Tris-HCl, pH 8.0, 150 mM NaCl and 2 mM DTT. The sequences of RNA oligonucleotides were identical to those used in EMSA. Data were collected by interference detection at 4 °C for all three protein concentrations (4 μM , 6 μM and 8 μM) at different rotor speeds (6,000, 8,500 and 12,000 r.p.m.). The buffer composition (density and viscosity) and protein partial specific volume (V -bar) were obtained using the SEDNTERP program (available through the Boston Biomedical Research Institute). The SE-AUC data were globally analysed using the Sedfit and Sedphat programs³⁷ and were fitted to a monomer–dimer equilibrium model to determine the dissociation constants (K_d) for the homodimers.

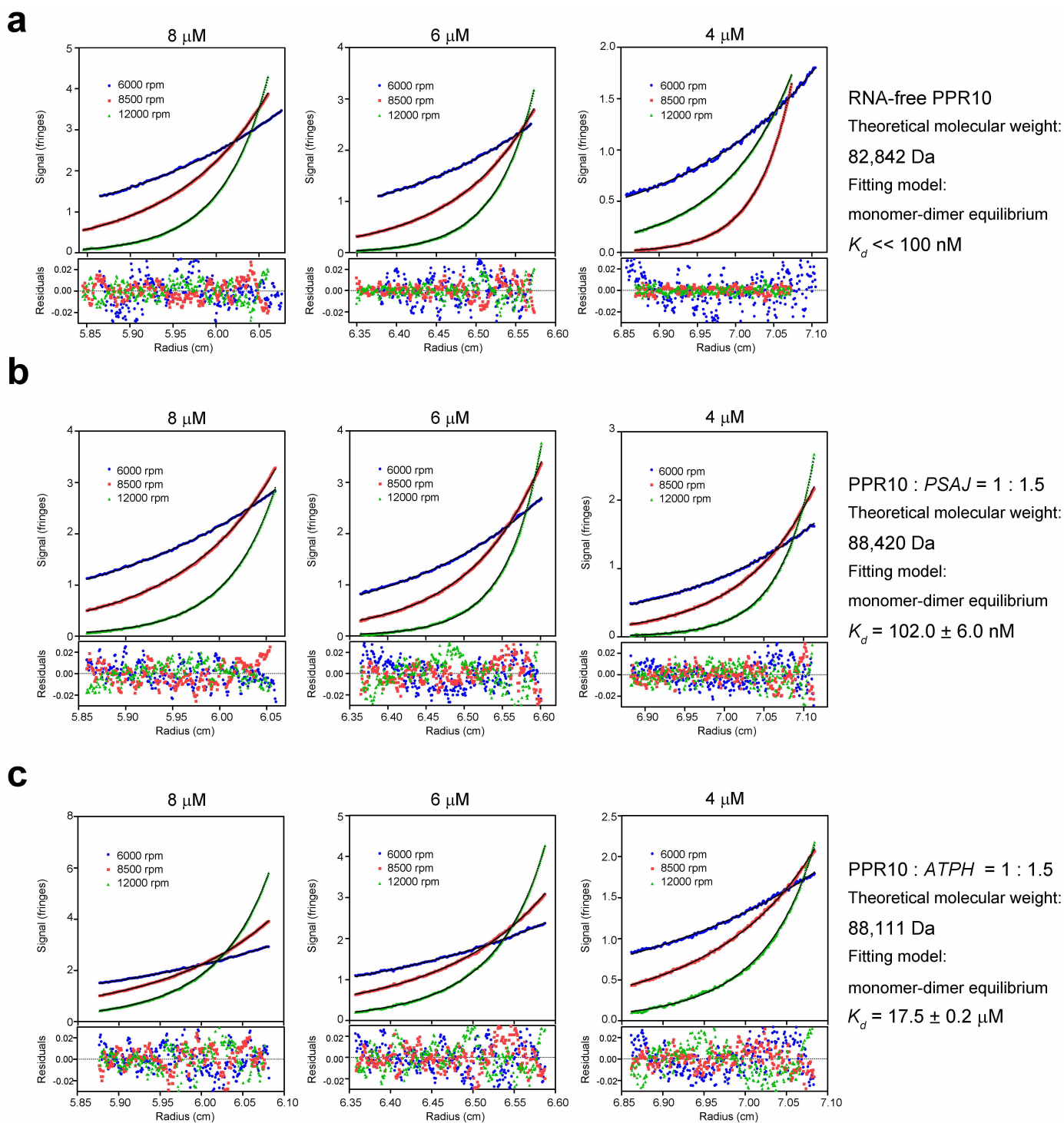
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58**, 1772–1779 (2002).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
- Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
- Schuck, P. On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. *Anal. Biochem.* **320**, 104–124 (2003).



Extended Data Figure 1 | Sequence alignment of the 19 repeats of PPR10.

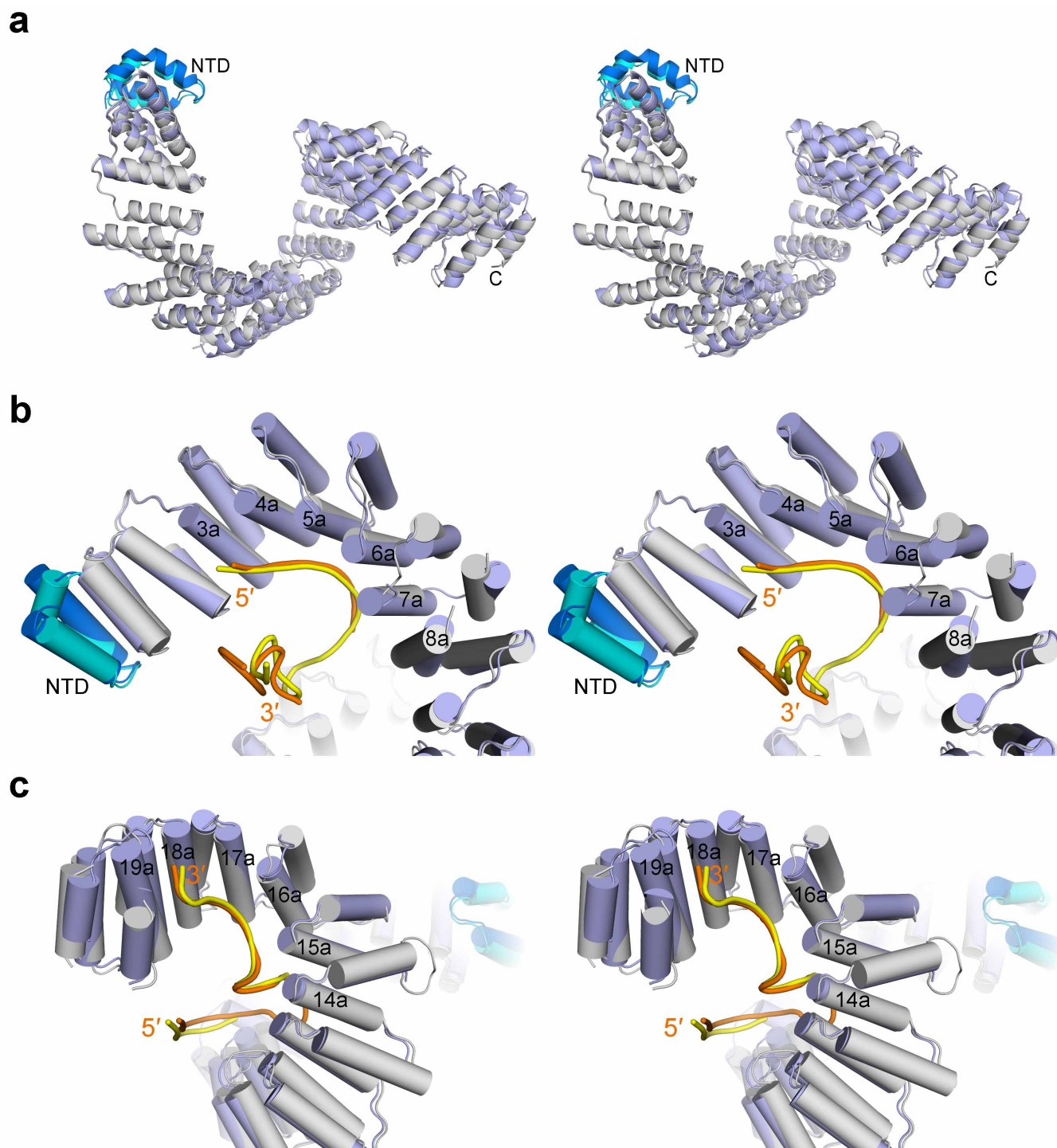
a, PPR10 from maize specifically recognizes two RNA elements. The cartoon above illustrates the predicted domain organization of PPR10. 1 and 19 refer to the repeat numbers. CTP, chloroplast transit peptide. The blue brick with '?' represents a fragment of approximately 30 amino acids whose function remains to be characterized. The minimal RNA elements of *PSAJ* and *ATPH* that are targeted by PPR10 are shown below the cartoon. **b**, Sequence alignment of 19 repeats in PPR10. The secondary structural elements of a typical PPR motif are shown above. The residues at the 2nd, 5th and 35th positions which were

predicted to be the molecular determinants for RNA-binding specificity are highlighted in magenta. The RNA sequences that can be recognized by PPR10 are listed on the right, 5' to 3' from top to bottom. The nucleotides which are recognized by PPR10 in a modular fashion in the *PSAJ*-PPR10 structure are shaded grey. **c**, The three numbering systems for a PPR motif. 1 is being used by Lurin *et al.*¹², Barkan *et al.*⁶ and others; 2 is adopted by the Pfam database and being used by Kobayashi *et al.*²⁸, Yagi *et al.*⁷ and others; and 3 is our proposed, structure-based numbering system. The residues that are predicted to specifically recognize RNA are coloured magenta.



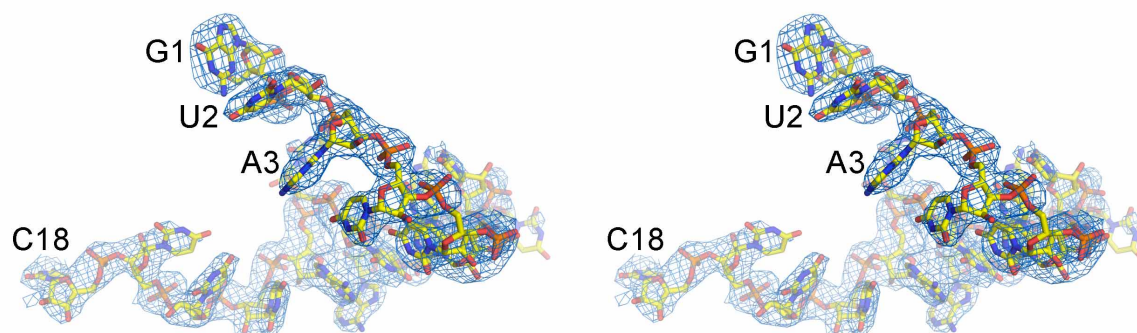
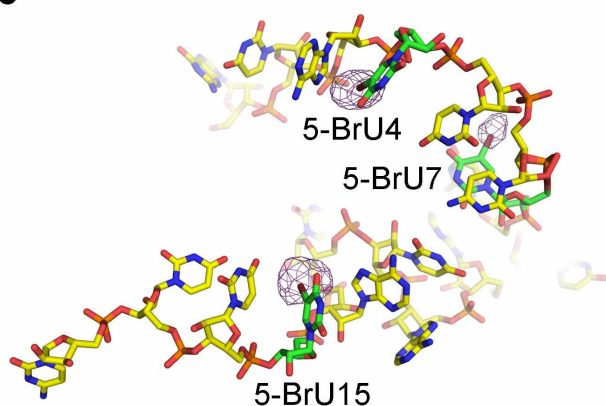
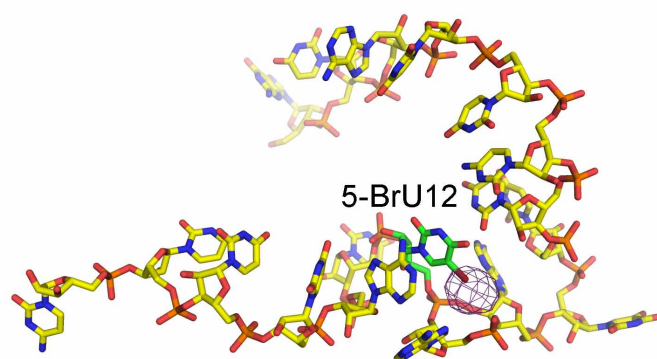
Extended Data Figure 2 | AUC-SE of PPR10 (residues 37–786, C256S/C279S/C430S/C449S) in the absence or presence of the target RNA elements. The molar concentrations of PPR10 are indicated above each panel.

PPR10 and the RNA oligonucleotides were mixed at a stoichiometric ratio of approximately 1:1.5. Details of the experiments are described in Methods.



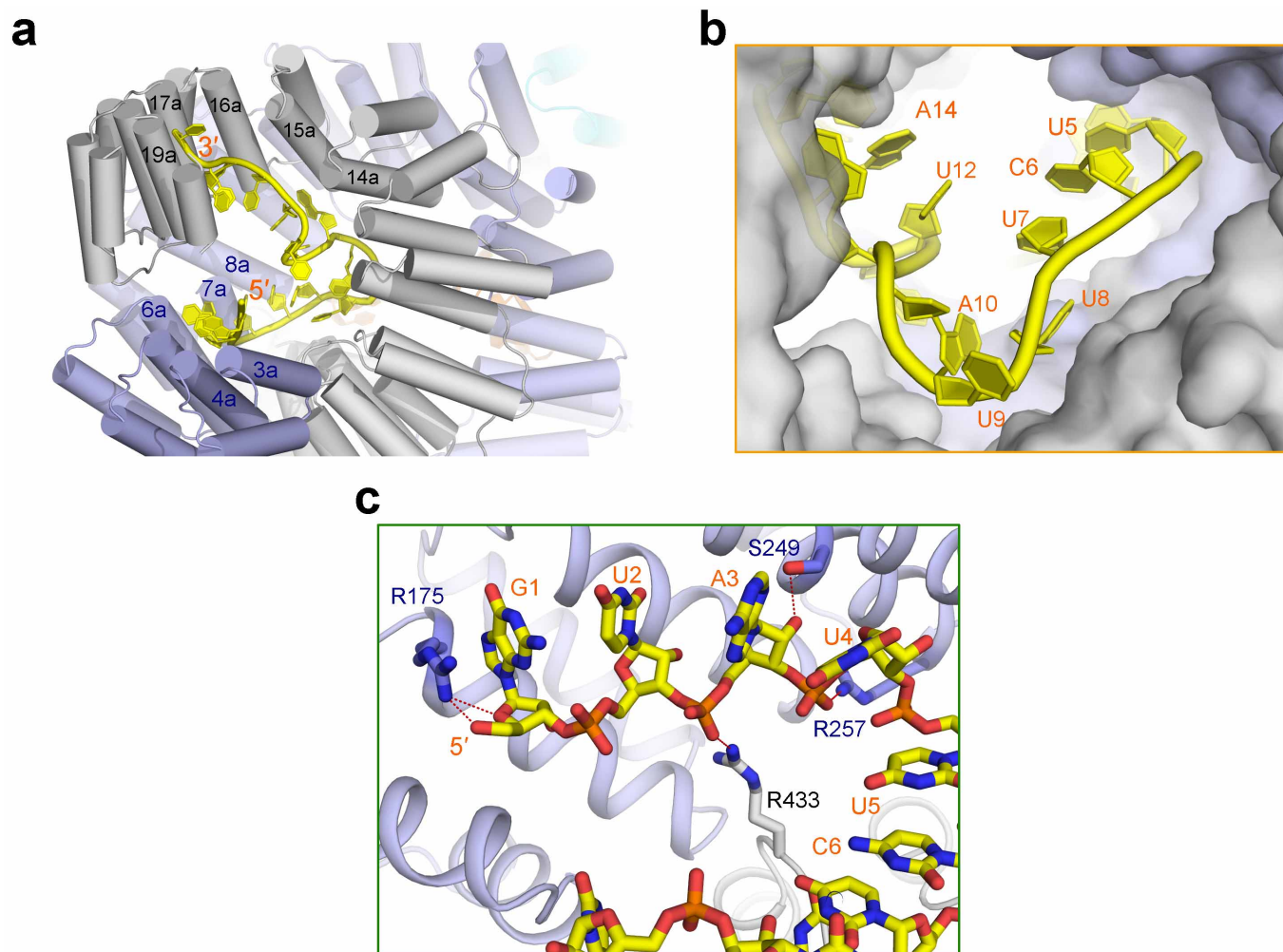
Extended Data Figure 3 | The two protomers of the RNA-bound PPR10 dimer exhibit similar conformations. **a**, The two protomers can be superimposed with a root-mean-squared deviation of 1.31 Å over 629 C α atoms. **b**, **c**, The two ssRNA segments are coordinated by the PPR10 dimer

similarly. The 5' and 3' segments of the bound PSAJ RNA are separately coordinated by the N-terminal repeats of one protomer (**b**), and the C-terminal repeats of the other protomer (**c**). Stereo-views are shown for all panels.

a**b****c**

Extended Data Figure 4 | Electron density maps for a bound ssRNA segment. **a**, The $2F_o - F_c$ electron density for one segment of the bound *PSA* RNA. The electron density, contoured at 1σ and coloured blue, is displayed in

stereo. **b**, **c**, The anomalous signals for bromine in the structures where the highlighted nucleotides were substituted with 5-bromouracil (5-BrU). The anomalous signals, shown in magenta mesh, are contoured at 5σ .

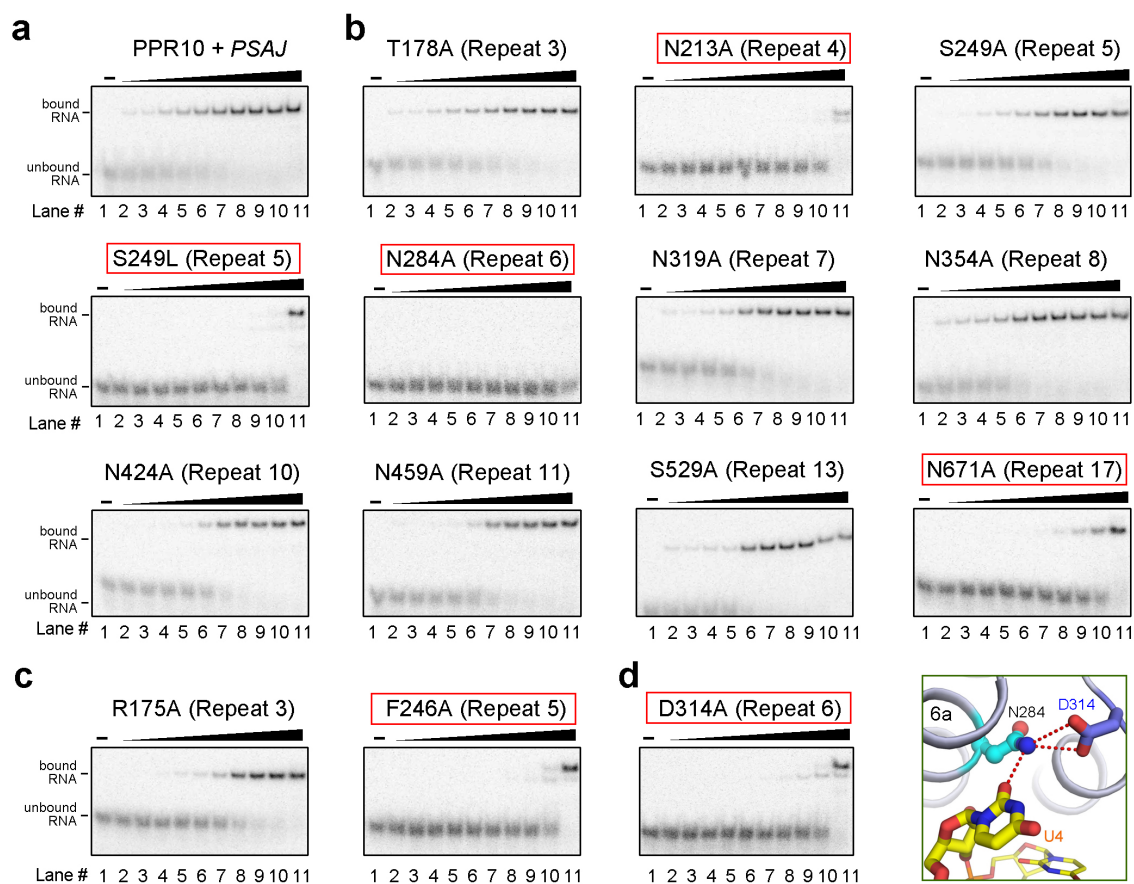


Extended Data Figure 5 | Coordination of the bound ssRNA by PPR10.

a, The 5' and 3' portions of the *PSAJ* RNA element are separately bound by the N-terminal and C-terminal repeats of the two PPR10 protomers. Shown here is a close-up view of the binding of *PSAJ* by one end of the PPR10 dimer.

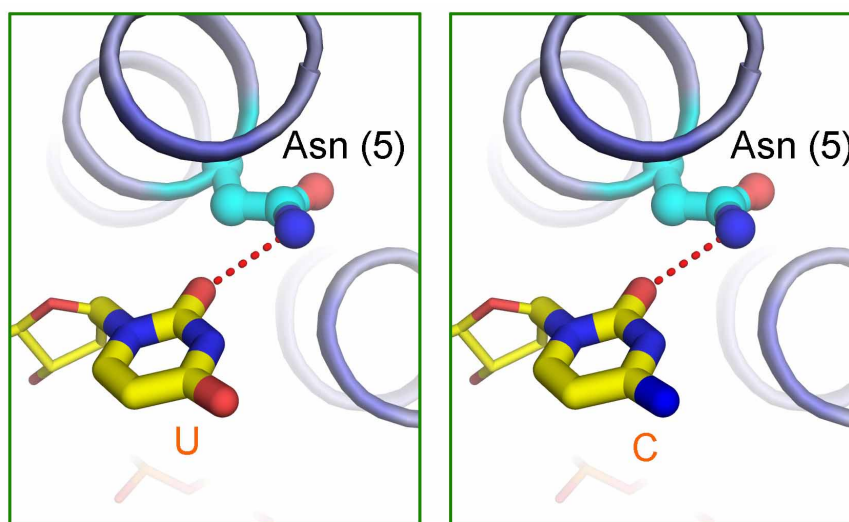
b, The nucleotides U5–A10, which form a U-turn in the ssRNA, are

uncoordinated in the cavity of the PPR10 dimer. The two protomers of PPR10 are shown in semi-transparent surface contour. **c,** The RNA backbone is coordinated by polar or charged residues through hydrogen bonds. The hydrogen bonds are represented by red dotted lines. The two protomers of PPR10 are coloured light purple and grey.



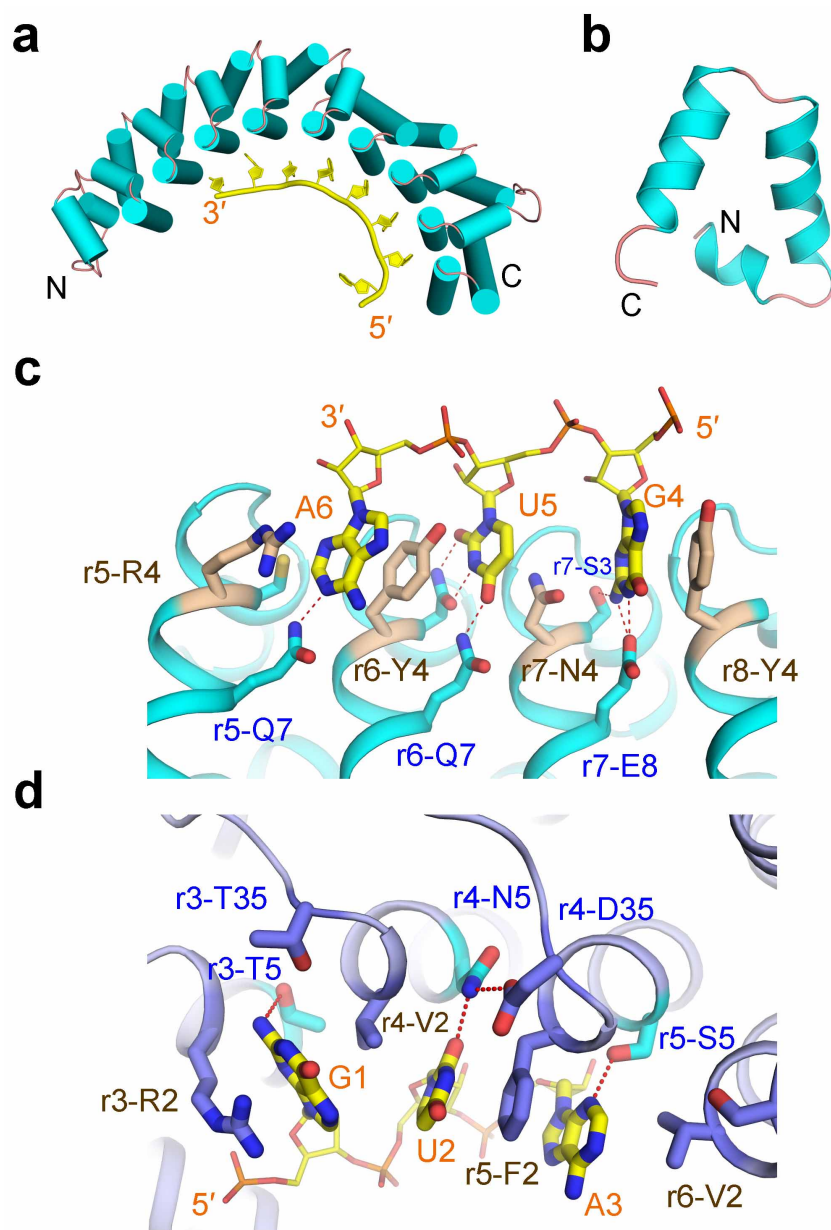
Extended Data Figure 6 | Mutational analysis of PPR10 residues that may be important for PSAJ. **a**, EMSA analysis of the interaction between PPR10 (residues 37–786, C256S/C279S/C430S/C449S) and PSAJ (5′-GUAUUCUUAAUUUUUC-3′). PPR10 was added with increasing concentrations of 0, 2, 4, 8, 16, 31, 63, 125, 250, 500, 1,000 nM in lanes 1–11 with approximately 40 pM ³²P-labelled PSAJ in each lane. **b**, Mutational analysis of

the 5th residues of the indicated PPR motifs. The indicated point mutations were introduced to PPR10 (residues 37–786, C256S/C279S/C430S/C449S). **c**, Examination of the 2nd residues in repeats 3 and 5. **d**, Examination of the 35th residue of repeat 6. Note that the side group of Asp 314 is hydrogen bonded to the side chain of Asn 284, the 5th residue of repeat 6. The same structural feature is also seen in repeat 4 (Fig. 3b).



Extended Data Figure 7 | The predicted coordination of base C by an Asn at the 5th position of a PPR motif. Left, the coordination of base U by Asn

observed in the structure. Right, the coordination of base C by Asn at the 5th position of a PPR motif modelled on the basis of the structure shown on the left.



Extended Data Figure 8 | Comparison of ssRNA coordination by PUF and PPR proteins. **a**, The structure of the human PUF protein PUM1 (also known as HSPUM) bound to the RNA element NRE1-19 (PDB accession code, 1M8W)²⁴. The PUF repeats constitute an arc with 8-nt ssRNA bound to the concave side. Notably, the orientations of the bound RNA and the protein are antiparallel, namely the 5' end is close to the C terminus of PUF. **b**, The structure of a PUF repeat. One canonical PUF repeat contains three helices, of which a short helix precedes a helical hairpin. **c**, Representative recognition of the RNA bases G, U, A by PUF repeats as seen in the structure of PUM1 bound to NRE1-19. The amino acids are labelled by the repeat number (r5, r6, r7, r8) followed by its one-letter code and position on the 2nd helix

within a PUF repeat (S3, N4, E8, and so on). The same scheme applies to **d**. **d**, The coordination of RNA bases G, U, A by PPR10. It is noteworthy that PUF and PPR proteins share several common features for RNA binding: (1) the ssRNA elements are coordinated by the helices on the inner layer; and (2) the base is sandwiched mostly by hydrophobic residues or Arg. Yet the differences are evident between the two families of repeat proteins. As seen in **c**, the RNA base is usually coordinated by two residues that are located at the 4th and the 7th positions on helix 2 within a PUF repeat. By contrast, the base is mainly coordinated by the 5th residue of a PPR motif. The 35th residue, the last residue of a PPR motif that is located at a loop region preceding the next PPR motif, also contributes to base recognition.

Extended Data Table 1 | Statistics of data collection and refinement. Values in parentheses are for the highest resolution shell

	Se-PPR10	PPR10-PSAJ complex
Number of crystals	1	5
Space Group	P2 ₁ 2 ₁ 2	P4 ₃
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	68.37, 176.54, 64.53	83.43, 83.43, 226.91
α , β , γ (°)	90, 90, 90	90, 90, 90
Wavelength (Å)	0.9793	0.9067
Resolution (Å)	40-2.85 (2.95-2.85)	40-2.45 (2.54-2.45)
R _{merge} (%)	12.7 (88.7)	8.9 (91.7)
I/sigma	28.9 (3.6)	27.7 (2.3)
Completeness (%)	100.0 (100.0)	96.3 (98.0)
Number of measured reflections	227,350	426,030
Number of unique reflections	18,683	54,087
Redundancy	12.2 (12.4)	7.9 (7.8)
Wilson B factor (Å ²)	73.3	82.9
R _{work} / R _{free} (%)	23.98/25.38	25.84/28.78
No. atoms		
Protein	5,326	10,384
main chain	2,808	5,480
side chain	2,518	4,904
RNA		652
water		114
B-factors		
Protein	74.8	85.45
main chain	75.2	85.13
side chain	74.3	85.81
RNA		93.84
water		67.51
R.m.s. deviations		
Bonds (Å)	0.011	0.015
Angle (°)	1.361	1.049
Ramachandran plot statistics (%)		
Most favorable	87.6	88.8
Additionally allowed	11.2	10.3
Generously allowed	1.1	0.9
Disallowed	0.0	0.0

Extended Data Table 2 | Statistics of data collection

	12-5BrU	4/7/15-5BrU	5/7/15-5BrU
Space Group	P4 ₃	P4 ₃	P4 ₃
Cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	83.48, 83.48, 227.94	83.11, 83.11, 227.68	83.30, 83.30, 226.10
α , β , γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Wavelength (Å)	0.9194	0.9194	0.9194
Resolution (Å)	40-3.2 (3.31-3.20)	40-3.00 (3.11-3.00)	40-3.20 (3.31-3.20)
R _{merge} (%)	12.2 (74.5)	11.7 (91.1)	14.8 (79.9)
I/sigma	21.2 (3.5)	26.2 (2.8)	20.1 (3.1)
Completeness (%)	100.0 (100.0)	99.8 (100.0)	100.0 (100.0)
Number of measured reflections	209,591	246,450	212,113
Number of unique reflections	25,708	30,609	25,347
Redundancy	8.2 (8.0)	8.1 (8.5)	8.4 (8.6)
Wilson B factor (Å ²)	69.0	79.6	71.4

Structure of LIMP-2 provides functional insights with implications for SR-BI and CD36

Dante Neculai¹, Michael Schwake^{2,3}, Mani Ravichandran⁴, Friederike Zunke², Richard F. Collins¹, Judith Peters², Mirela Neculai⁴, Jonathan Plumb¹, Peter Loppnau⁴, Juan Carlos Pizarro^{4,5}, Alma Seitova⁴, William S. Trimble^{1,6}, Paul Saftig², Sergio Grinstein^{1,6,7} & Sirano Dhe-Paganon^{4,8}

Members of the CD36 superfamily of scavenger receptor proteins are important regulators of lipid metabolism and innate immunity. They recognize normal and modified lipoproteins, as well as pathogen-associated molecular patterns. The family consists of three members: SR-BI (which delivers cholesterol to the liver and steroidogenic organs and is a co-receptor for hepatitis C virus), LIMP-2/LGP85 (which mediates lysosomal delivery of β -glucocerebrosidase and serves as a receptor for enterovirus 71 and coxsackieviruses) and CD36 (a fatty-acid transporter and receptor for phagocytosis of effete cells and *Plasmodium*-infected erythrocytes). Notably, CD36 is also a receptor for modified lipoproteins and β -amyloid, and has been implicated in the pathogenesis of atherosclerosis and of Alzheimer's disease¹. Despite their prominent roles in health and disease, understanding the function and abnormalities of the CD36 family members has been hampered by the paucity of information about their structure. Here we determine the crystal structure of LIMP-2 and infer, by homology modelling, the structure of SR-BI and CD36. LIMP-2 shows a helical bundle where β -glucocerebrosidase binds, and where ligands are most likely to bind to SR-BI and CD36. Remarkably, the crystal structure also shows the existence of a large cavity that traverses the entire length of the molecule. Mutagenesis of SR-BI indicates that the cavity serves as a tunnel through which cholesterol(esters) are delivered from the bound lipoprotein to the outer leaflet of the plasma membrane. We provide evidence supporting a model² whereby lipidic constituents of the ligands attached to the receptor surface are handed off to the membrane through the tunnel, accounting for the selective lipid transfer characteristic of SR-BI and CD36.

The ectodomain of human LIMP-2 comprises a new fold with an antiparallel β -barrel core with many short α -helical segments (Fig. 1a and Supplementary Figs 1 and 2). The human LIMP-2 β -barrel is imperfect, with one of the 13 slats being a short loop (amino acids 252–255) that partly fills the centre of the barrel. The barrel is asymmetric, with four β -strands extending for the entire length of the human LIMP-2 structure, forming a prolate spheroid. The β -barrel and its extended β -strands are topped at the head by a three-helix bundle (Fig. 1 and Supplementary Fig. 2). The α 1 and α 15 helices belong to the amino (N-) and carboxy (C-)terminal secondary structure elements and define the boundaries of the ectodomain (Fig. 1). The electron density map of human LIMP-2 shows which residues are part of the core globular domain (Supplementary Fig. 2) and therefore cannot function as linkers. Because the transmembrane segments are well defined (Supplementary Fig. 1), we conclude that at most six residues may act as linkers between the globular core and the transmembrane segments. Even if stretched, these residues would afford only a distance of a few ångströms between the surface of the membrane and the globular domain of the protein. Hence, the ectodomain

of LIMP-2 (and in all likelihood also those of SR-BI and CD36) must lie very close or be directly apposed to the exofacial leaflet of the membrane.

Two disulphide bridges stabilize the fold (Fig. 1 and Supplementary Figs 1, 2 and 4). The buried C274–C329 bridge links the α 9– β 11 and β 13– β 14 loops; the partly solvent-exposed C312–C318 bridge stabilizes the α 11– β 13 loop, and the lone reduced cysteine, C245, is partly protected from solvent by a long glycosylation chain at N249 (Fig. 1 and Supplementary Fig. 4). The disulphide bridge pattern of LIMP-2 is similar to the predicted SR-BI (C321–C323, C274–C329) and CD36 (C313–C322, C272–C333) disulphide bridges, according to structural homology models (Supplementary Figs 1 and 4). The modelled SR-BI and CD36 disulphide bridges were consistent with those deduced experimentally^{3,4} (Supplementary Fig. 4). The crystallization construct included only nine out of the ten glycosylation sequons of human LIMP-2 (Supplementary Figs 1 and 3), and well-defined electron densities were observed for all nine glycosylation sites (Supplementary Fig. 2). N-acetyl-hexosamine chains are distributed equally around the lower half of the domain (Fig. 1) on asparagine residues 45, 68, 105, 206, 224, 249, 304, 325 and 412. In the crystal structure, the carbohydrate chain length varies from one to five hexose units. We used mutagenesis to analyse the role of carbohydrate moieties in directing the localization of mouse LIMP-2 (Supplementary Fig. 5). Like the wild type, 9 of the 11 glycosylation site mutants targeted to lysosomes (Supplementary Fig. 5), whereas N68Q and N325Q were retained in the endoplasmic reticulum (Supplementary Fig. 5).

Although glycosylation sites are evenly distributed around the mid-section of the protein, the head is free of any glycosylation sites and is composed of a three-helix bundle formed by α -helices 4, 5 and 7, and two other short abutting helices, α 2 and α 14, and the β 7 strand (Fig. 1). The solvent-exposed side chains of the α 2 and α 14 helices and of the β 7 strand are predominantly charged and form the apex of the human LIMP-2 structure, whereas the three-helix bundle, particularly the area between α 5 and α 7, projects hydrophobic side chains (the helix bundle face), suggesting a possible interaction site for ligands or for dimerization (Fig. 1 and Supplementary Fig. 6). Indeed, the crystal structure shows close apposition of the hydrophobic helices between human LIMP-2 monomers (Supplementary Fig. 6).

LIMP-2 was described as a chaperone for β -glucocerebrosidase (β -GCCase), with which it forms a tight complex both *in vitro* and *in vivo* (Fig. 2). The interaction site in LIMP-2 was previously identified as region 150–167 (ref. 5). From the human LIMP-2 crystal structure, the boundaries of helix α 5 can be redrawn to 149–163. The region of residues 164–167 represents a loop preceding the β 7 strand. Extending previous studies^{5,6}, we found that point mutations not only in α 5 but also in α 7 showed impaired β -GCCase binding to mouse LIMP-2 (Fig. 2 and Supplementary Fig. 7). Interestingly, a point mutation

¹Cell Biology Program, The Hospital for Sick Children, Toronto M5G 1X8, Canada. ²Biochemisches Institut der Christian-Albrechts-Universität zu Kiel, Otto-Hahn-Platz 9, 24118 Kiel, Germany. ³Biochemie III, Fakultät für Chemie, Universität Bielefeld, Universitätsstrasse 25, 33615 Bielefeld, Germany. ⁴Structural Genomics Consortium, University of Toronto, 101 College Street, Toronto, Ontario M5G 1L7, Canada. ⁵Department of Tropical Medicine, Tulane University, New Orleans, Louisiana 70118, USA. ⁶Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada. ⁷Keenan Research Centre of the Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, Toronto M5C 1N8, Canada. ⁸Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

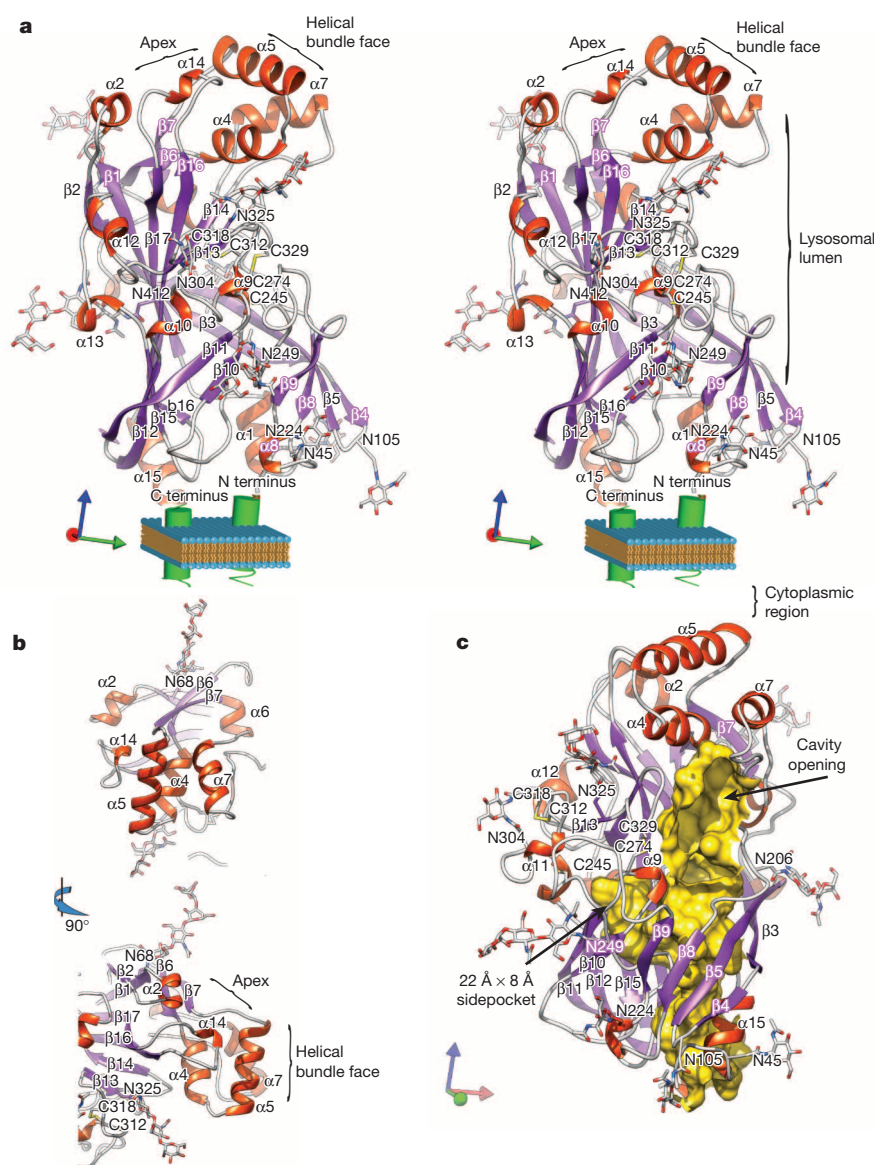
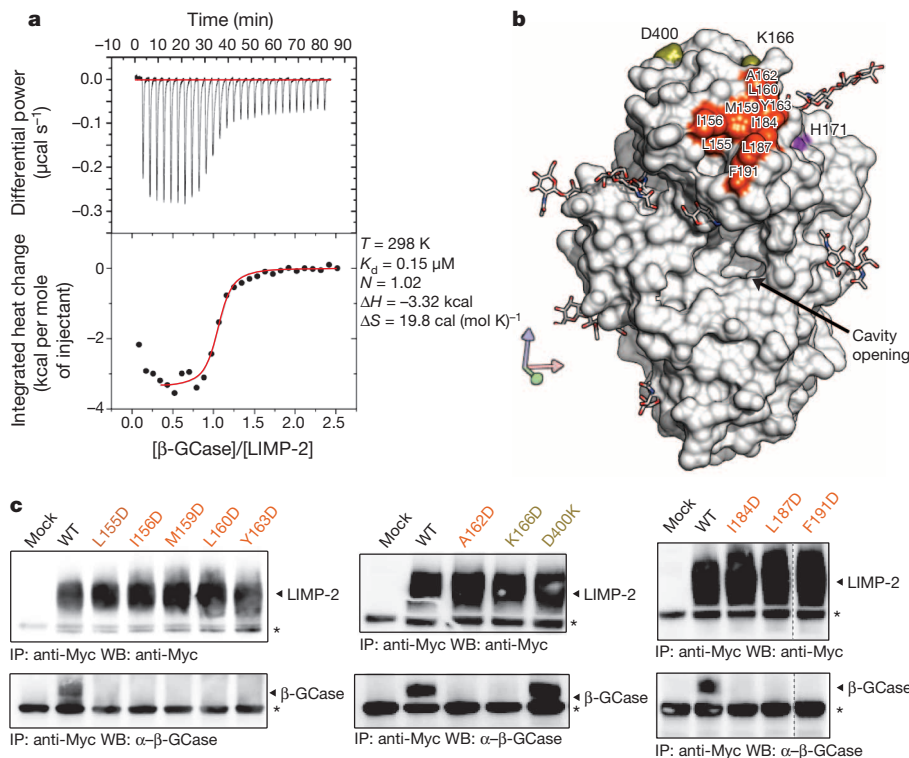


Figure 1 | Structure of the human LIMP-2 luminal domain. **a**, Stereo view of human LIMP-2 crystal structure, emphasizing its likely orientation with respect to the bilayer. The α -helices and β -strands are shown in labelled ribbon format in orange and purple, respectively. Also included are glycosylation sites, cysteine residues and disulphide bonds (yellow) in stick format. Transmembrane segments and the lipid bilayer are schematically illustrated. Reference coordinates are shown (and maintained in subsequent figures), in which the plane formed by the green and red arrows is parallel with the membrane, and the blue arrow is perpendicular to the membrane. **b**, Ribbon representations of the human LIMP-2 helical bundle in two orientations, rotated by 90° relative to each other. **c**, Representation of the human LIMP-2 luminal tunnel, shown as a yellow surface.

(K166D)—located on the apex of LIMP-2—greatly impaired β -GCCase binding. In close proximity to residue K166—and also located on the apex—is H171 (Fig. 2), which was shown to function in the pH-dependent control of β -GCCase release in the lysosome⁷. Graphical representations showed that the electrostatic potential of the apex of human LIMP-2 differs markedly from those of SR-BI or CD36 (Supplementary Fig. 4). In contrast to human LIMP-2, the apex regions of both human SR-BI and human CD36 show a remarkable accumulation of cationic residues (Supplementary Fig. 4). An electrostatic interaction at the apex probably contributes to the association of these receptors with their polyanionic ligands^{8,9}. It is also noteworthy that several SNPs, and other residues previously identified as involved in ligand binding by CD36 family members^{10–13}, map to the apex region (Supplementary Fig. 8). To explore further the role of the apex region and helical bundle face in ligand binding we performed mutagenesis experiments. As illustrated in Fig. 3 and Supplementary Figs 9–11, mutations on either the apex or the face of the helical bundle of SR-BI precluded binding of DiI-labelled high-density lipoprotein (HDL), oxidized low-density lipoprotein (oxLDL) and acetylated LDL. We also studied human CD36. Mutations on its apex or helical bundle abrogated binding of DiI-labelled oxLDL (Fig. 3 and Supplementary Figs 9 and 10). Together, these findings—which were validated by co-immunoprecipitation (for LIMP-2 see Fig. 2 and Supplementary

Fig. 7), immunofluorescence (for SR-BI, CD36 see Fig. 3 and Supplementary Figs 9 and 11) and flow cytometry (for SR-BI, CD36 see Fig. 3 and Supplementary Fig. 10)—support the notion that the helical bundle face and the apex region that forms the head of the CD36 superfamily members are important for ligand binding.

Another striking feature of the structure of LIMP-2, which is recapitulated in models of SR-BI and CD36, is the presence of interconnected cavities that form a tunnel through the entire length of the ectodomain (Figs 1 and 4 and Supplementary Fig. 12). In the case of LIMP-2, there is a solvent-exposed cavity between the $\alpha 12$ – $\beta 12$ loop and the $\alpha 7$ – $\beta 7$ loop with a 5 Å × 5 Å opening (Supplementary Fig. 12). This deep ridge joins with the most prominent cavity, an approximately 22 Å × 11 Å × 8 Å (length × width × height; Supplementary Fig. 12) pocket, at the centre of the β -barrel. Finally, the channel emerges at the tail end of the domain, bounded by the N- and C-terminal helices, towards the exofacial (luminal in LIMP-2) leaflet of the membrane. The cavities are predominantly, but not



entrance, do not seem to form ionic interactions (Supplementary Fig. 12).

SR-BI is proposed to mediate the delivery of cholesterol(esters) from HDL^{2,14–16} to the membrane and vice versa^{17,18}. The selective nature of this transport—lipids are taken up by the cells in excess over the

apolipoproteins—prompted Rodriguez *et al.*² to suggest that a non-aqueous ‘channel’ may exist whereby cholesterol(esters) are transferred to the membrane. As illustrated in Supplementary Fig. 2, the width of the LIMP-2 tunnel is sufficient to accommodate a polyethyleneglycol molecule (from the crystallization buffer) and we speculate

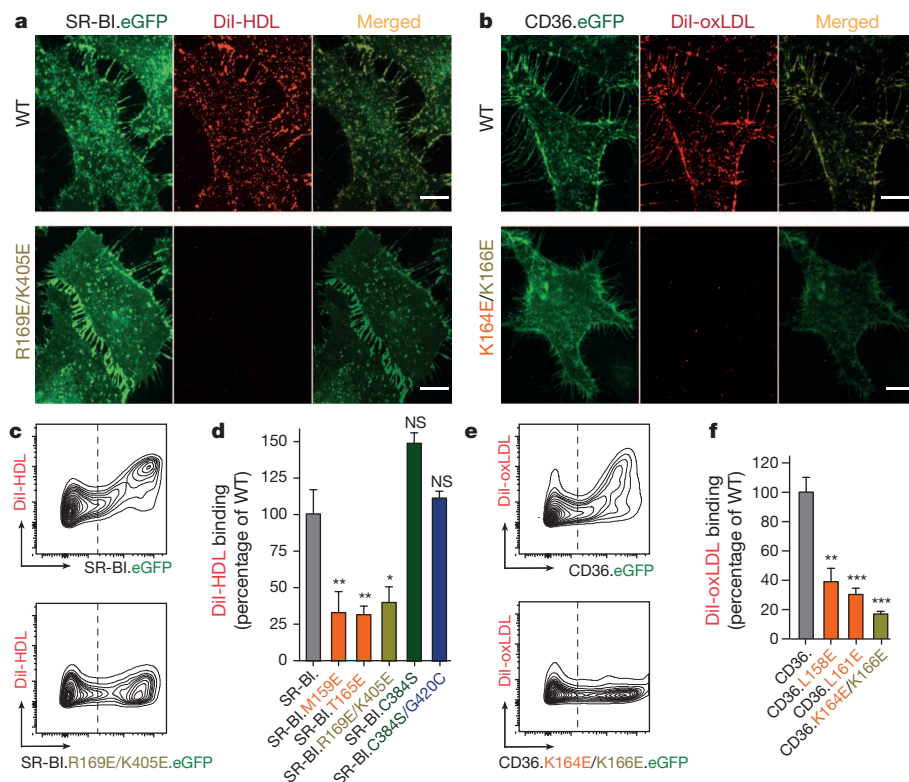


Figure 3 | SR-BI and CD36 use the same interface as LIMP-2 for ligand binding. **a**, HeLa cells expressing wild-type or mutant human SR-BI (*hSR-BI.R169E/K405E*) fused to enhanced green fluorescent protein (eGFP) at the C terminus were incubated with DiI-labelled HDL for 20 min at 10 $^{\circ}$ C, followed by examination by spinning-disc confocal microscopy. Scale bars, 10 μ m. **b**, HeLa cells expressing wild-type or mutant human CD36 (*hCD36.K164E/K166E*) fused to eGFP were incubated with DiI-labelled oxLDL for 20 min at 10 $^{\circ}$ C followed by microscopic examination. **c**, **e**, CHO cells expressing wild-type or mutant human SR-BI and human CD36 fused to eGFP were incubated with DiI-labelled HDL and DiI-labelled oxLDL for 20 min at 10 $^{\circ}$ C followed by flow cytometry. Representative results for DiI-labelled HDL or DiI-labelled oxLDL binding to wild-type or mutant human SR-BI (*hSR-BI.R169E/K405E*) or wild type and mutant human CD36 (*hCD36.K164E/K166E*) tagged with eGFP. **d**, **f**, Quantification of DiI-labelled HDL or DiI-labelled oxLDL binding to CHO cells transiently transfected with wild-type and mutant human SR-BI and human CD36, respectively. All values were normalized to correct for differences in receptor surface expression relative to wild-type SR-BI or wild-type CD36, based on flow cytometry (Supplementary Fig. 10). Statistical analyses were performed using one-way analysis of variance with Tukey post-testing using GraphPad Prism version 5. * P < 0.05, ** P < 0.01, *** P < 0.001 compared with wild type. Error bars, s.e.m.

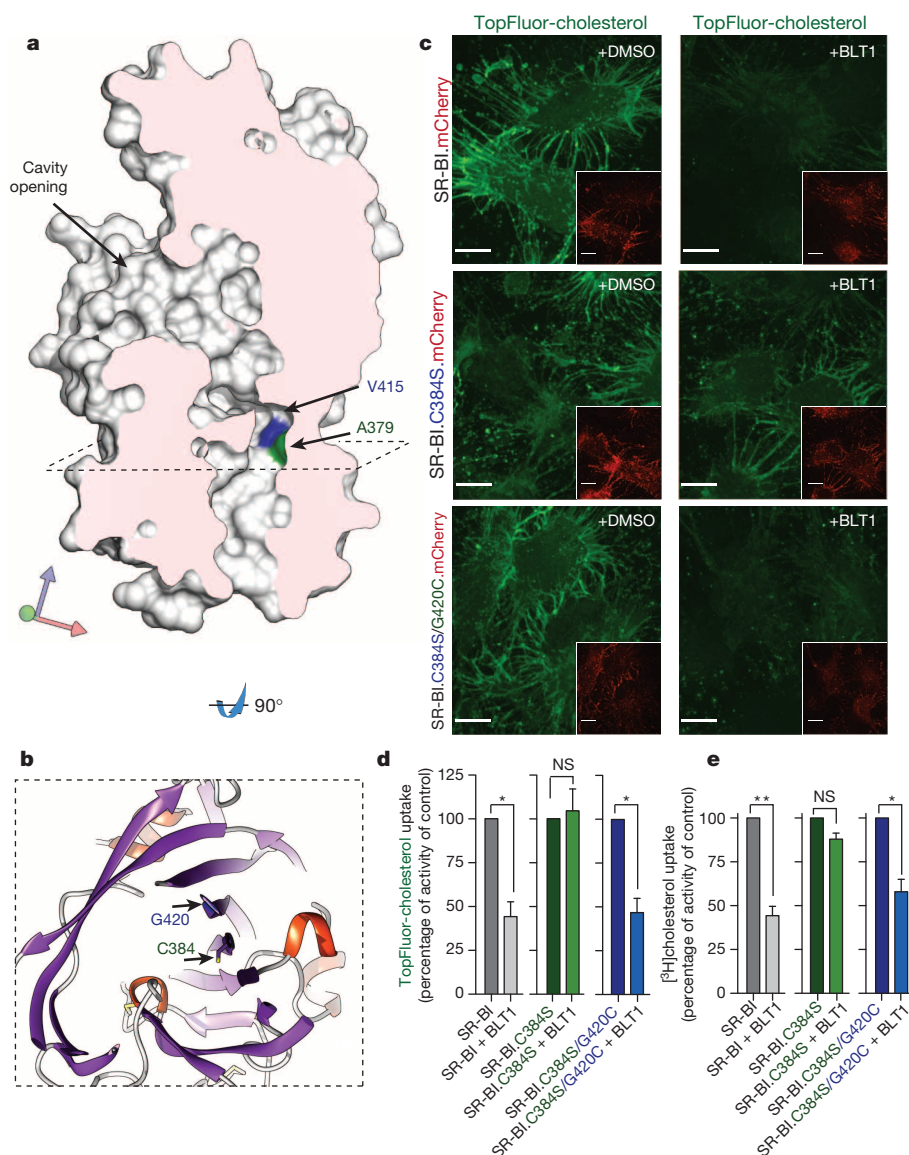


Figure 4 | SR-BI has a functionally important tunnel. **a**, Representation of the human LIMP-2 luminal tunnel, shown as a slice through the body of the protein (sliced solid shown in pink). Residues A379 and V415, which point towards the cavity of the tunnel, are shown in green and blue, respectively. **b**, Close-up view of C384 and G420 of the human SR-BI structural homologue (dashed square in **b** is equivalent to the area indicated by the dashed plane in **a**). **c**, Transport of TopFluor-cholesterol from reconstituted HDL in HeLa cells. HeLa cells were transiently transfected with wild-type and mutant forms (SR-BI.C384S; SR-BI.C384S/G420C) of mCherry-tagged SR-BI. The cells were pre-incubated in ATP-depletion buffer and, where indicated, with BLT1 (right panel) for 30 min at 37 °C. Next, cells were exposed to labelled HDL for 10 min at 37 °C, washed and chased for 40 min at 37 °C. Finally, a 50-fold excess HDL was added. Insets represent distribution of wild-type and mutant forms of mCherry-tagged SR-BI. Scale bars, 10 μm. **d**, Quantification of TopFluor-cholesterol uptake by HeLa cells transfected with wild-type or mutant forms of SR-BI with or without BLT1. The ratio of green (TopFluor-cholesterol) over red (SR-BI fused to mCherry) fluorescence was used to estimate uptake; data were normalized to control. Averages of three independent experiments. * $P < 0.05$, NS, not significant. **e**, Uptake of [³H]cholesterol by CHO cells transfected with wild-type or mutant SR-BI. Cells were pre-incubated in serum-free DMEM and, where indicated, with BLT1 for 60 min at 37 °C. Next, cells were exposed to sub-saturating amounts of HDL doubly labelled with [³H]cholesterol and TRITC for 2 h at 37 °C, washed and collected. Total red and green fluorescence was measured, followed by measurement of [³H]cholesterol radioactivity. The resulting ratios were normalized to the untreated control. Statistical analyses comparing samples with or without BLT1 were performed using a paired two-tailed *t*-test at 95% confidence intervals using GraphPad Prism version 5. * $P < 0.05$, ** $P < 0.01$ compared with control (dimethylsulphoxide only). Error bars, s.e.m.

that the equivalent tunnel of SR-BI can accommodate cholesterol(esters). This possibility is given credence by the observations of Krieger and colleagues³, who reported that cholesterol(ester) uptake by SR-BI is inhibited in a C384-dependent manner by a thiosemicarbazone reagent, blocker of lipid transport 1 (BLT1)^{19,20}. C384 is located in the lumen of the SR-BI tunnel and, based on earlier results³, the attachment of the bulky BLT1 is expected to block the transit of cholesterol through the tunnel (Fig. 4). Similarly, mutation of C384 in SR-BI with residues that have larger side chains than cysteine (L, Y) phenocopied the effects of treatment with BLT1, whereas mutation of C384 for smaller residues (G, A) produced no such phenotype²¹. Of note, exposure of the other free cysteine residue on SR-BI (C251; Supplementary Fig. 3) to BLT1 was without effect¹⁸, in all likelihood because C251 resides on the outer surface of the molecule, away from the tunnel (Fig. 1 and Supplementary Fig. 12). Fluorescence and radioactivity uptake experiments were used to validate the existence of a lipid-translocating tunnel in SR-BI, taking advantage of the observation that the effect of BLT1 is lost when C384 is converted to serine³. We generated a double mutant C384S/G420C, resulting in the *de novo* introduction of a reactive cysteine residue in the lining of the tunnel. Although the mutation itself had little effect, treatment of the C384S/G420C double mutant receptors with BLT1 produced an inhibition of transport comparable to that induced by this compound in the wild-type receptor (Fig. 4).

In summary, we establish the structure and folding of members of the CD36 receptor superfamily. The exofacial (extracellular or luminal, depending on the receptor; Supplementary Fig. 1) domain shows two salient features: a helical bundle where ligands bind, and a tunnel seemingly used for delivery of lipid substrates to and from the membrane. The electrostatic vector of SR-BI and CD36 would ensure firm retention of the apoprotein-containing ligand while lipids are discharged. Because the best-known function of LIMP-2 is to mediate the delivery of β -GCase from the endoplasmic reticulum to lysosomes, the requirement for a hydrophobic tunnel in this case remains unexplained. It is conceivable that, having released β -GCase to the lumen of lysosomes, LIMP-2 can serve a second function involving lipid transport. Delivery of saposin-associated lipids or the products of their hydrolysis to the membrane of the lysosome by LIMP-2 is an attractive possibility that needs to be tested experimentally. More importantly, we provide a plausible mechanism for the selective lipid uptake by SR-BI and CD36, whereby their exofacial domains form a channel through which lipids can be translocated to the membrane bilayer.

METHODS SUMMARY

An Sf9 insect cell melittin signal-peptide secretion system was used to express the ectodomain of human LIMP-2 (amino acids 35–430; Supplementary Fig. 1); immobilized metal-ion affinity chromatography and size-exclusion chromatography

were used for purification. Although untreated human LIMP-2 crystals diffracted poorly, extensive dehydration followed by X-ray diffraction screening produced a 3 Å dataset. Phases were obtained by single anomalous dispersion collected from crystals soaked with HgCl₂ and Ta₆Br₁₄. Six molecules were packed in the asymmetric unit (four molecules forming two head-to-head dimers and two monomeric molecules) with a root mean squared deviation across all atoms of 1.2 Å, producing a 3 Å resolution model with excellent statistics (Supplementary Table 1 and Supplementary Fig. 2). Details of this and other methods can be found in the Supplementary Information.

Received 9 December 2012; accepted 20 September 2013.

Published online 27 October 2013.

- Canton, J., Neculai, D. & Grinstein, S. Scavenger receptors in homeostasis and immunity. *Nature Rev. Immunol.* **13**, 621–634 (2013).
- Rodriguez, W. V. et al. Mechanism of scavenger receptor class B type I-mediated selective uptake of cholesteryl esters from high density lipoprotein to adrenal cells. *J. Biol. Chem.* **274**, 20344–20350 (1999).
- Yu, M. et al. Exoplasmic cysteine Cys384 of the HDL receptor SR-BI is critical for its sensitivity to a small-molecule inhibitor and normal lipid transport activity. *Proc. Natl Acad. Sci. USA* **108**, 12243–12248 (2011).
- Rasmussen, J. T., Berglund, L., Rasmussen, M. S. & Petersen, T. E. Assignment of disulfide bridges in bovine CD36. *Eur. J. Biochem.* **257**, 488–494 (1998).
- Reczek, D. et al. LIMP-2 is a receptor for lysosomal mannose-6-phosphate-independent targeting of β-glucocerebrosidase. *Cell* **131**, 770–783 (2007).
- Blanz, J. et al. Disease-causing mutations within the lysosomal integral membrane protein type 2 (LIMP-2) reveal the nature of binding to its ligand β-glucocerebrosidase. *Hum. Mol. Genet.* **19**, 563–572 (2010).
- Zachos, C., Blanz, J., Saftig, P. & Schwake, M. A critical histidine residue within LIMP-2 mediates pH sensitive binding to its ligand β-glucocerebrosidase. *Traffic* **13**, 1113–1123 (2012).
- Ryeom, S. W., Silverstein, R. L., Scotto, A. & Sparrow, J. R. Binding of anionic phospholipids to retinal pigment epithelium may be mediated by the scavenger receptor CD36. *J. Biol. Chem.* **271**, 20536–20539 (1996).
- Jimenez-Dalmaroni, M. J. et al. Soluble CD36 ectodomain binds negatively charged diacylglycerol ligands and acts as a co-receptor for TLR2. *PLoS ONE* **4**, e7411 (2009).
- Puente Navazo, M. D., Daviet, L., Ninio, E. & McGregor, J. L. Identification on human CD36 of a domain (155–183) implicated in binding oxidized low-density lipoproteins (Ox-LDL). *Arterioscler. Thromb. Vasc. Biol.* **16**, 1033–1039 (1996).
- Stylanou, I. M. et al. Novel ENU-induced point mutation in scavenger receptor class B, member 1, results in liver specific loss of SCARB1 protein. *PLoS ONE* **4**, e6521 (2009).
- Chadwick, A. C. & Sahoo, D. Functional characterization of newly-discovered mutations in human SR-BI. *PLoS ONE* **7**, e45660 (2012).
- Kar, N. S., Ashraf, M. Z., Valiyaveetil, M. & Podrez, E. A. Mapping and characterization of the binding site for specific oxidized phospholipids and oxidized low density lipoprotein of scavenger receptor CD36. *J. Biol. Chem.* **283**, 8765–8771 (2008).
- Rigotti, A., Miettinen, H. E. & Krieger, M. The role of the high-density lipoprotein receptor SR-BI in the lipid metabolism of endocrine and other tissues. *Endocr. Rev.* **24**, 357–387 (2003).
- Gu, X. et al. The efficient cellular uptake of high density lipoprotein lipids via scavenger receptor class B type I requires not only receptor-mediated surface binding but also receptor-specific lipid transfer mediated by its extracellular domain. *J. Biol. Chem.* **273**, 26338–26348 (1998).
- Liu, B. & Krieger, M. Highly purified scavenger receptor class B, type I reconstituted into phosphatidylcholine/cholesterol liposomes mediates high affinity high density lipoprotein binding and selective lipid uptake. *J. Biol. Chem.* **277**, 34125–34135 (2002).
- Gu, X., Kozarsky, K. & Krieger, M. Scavenger receptor class B, type I-mediated [³H]cholesterol efflux to high and low density lipoproteins is dependent on lipoprotein binding to the receptor. *J. Biol. Chem.* **275**, 29993–30001 (2000).
- Ji, Y. et al. Scavenger receptor BI promotes high density lipoprotein-mediated cellular cholesterol efflux. *J. Biol. Chem.* **272**, 20982–20985 (1997).
- Nieland, T. J., Penman, M., Dori, L., Krieger, M. & Kirchhausen, T. Discovery of chemical inhibitors of the selective transfer of lipids mediated by the HDL receptor SR-BI. *Proc. Natl Acad. Sci. USA* **99**, 15422–15427 (2002).
- Nieland, T. J. et al. Identification of the molecular target of small molecule inhibitors of HDL receptor SR-BI activity. *Biochemistry* **47**, 460–472 (2008).
- Yu, M., Lau, T. Y., Carr, S. A. & Krieger, M. Contributions of a disulfide bond and a reduced cysteine side chain to the intrinsic activity of the high-density lipoprotein receptor SR-BI. *Biochemistry* **51**, 10044–10055 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank W. Temple, D. Cossar, S. Graslund, C. Arrowsmith, R. Stahelin, L. Andresen, J. Groth, M. Langer and J. Scott for assistance and discussions. This study was supported by the Canadian Institutes for Health Research (grants MOP-102474 and MOP-126069) and the Deutsche Forschungsgemeinschaft (grants GRK1459 to M.S. and SFB877). F.Z. is supported through the Böhlinger Ingelheim Fonds. W.S.T. is the recipient of a Canada Research Chair in Molecular Cell Biology. The Structural Genomics Consortium is a registered charity (number 1097737) that receives funds from the Canadian Institutes for Health Research, the Canada Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck & Co., the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research, and the Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions S.G., W.S.T., P.S., M.S., S.D.-P. and D.N. designed the research; D.N., M.R., M.N., P.L., A.S., J.C.P., R.C., J. Plumb, F.Z. and J. Peters performed the experiments; S.G., W.S.T., P.S., M.S., S.D.-P., D.N. and F.Z. analysed the results; S.G. wrote the paper; W.S.T., P.S., M.S., S.D.-P., D.N., J. Peters and F.Z. edited and commented on the manuscript.

Author Information Atomic coordinates and structure factors have been deposited in Protein Data Bank under accession number 4F7B. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.G. (sergio.grinstein@sickkids.ca).

CAREERS

RELOCATION Mobile scientists do better than those who stay in their home nations **p.179**

FLORIDA All-star scientists sought as part of university plan **p.179**

NATUREJOBS For the latest career listings and advice www.naturejobs.com

IKON IMAGES/CORBIS



BY SARAH WEBB

John Long never expected to write a book. His weeks were crammed with biology research, writing papers and teaching at Vassar College in Poughkeepsie, New York. But in 2009, after multiple newspapers picked up an Associated Press story about his work using robots to study evolution, he got a call from a literary agent in New York City.

The agent encouraged him to write a popular, mainstream book for general sale. Long was sceptical that such a project was worth his time or would advance his career, but the agent persisted. “You’re a teacher, right?” she asked. She convinced him that a book would allow him to reach “fun-loving nerds who want to learn something cool about science”, he recalls. Basic Books published *Darwin’s Devices* in 2012; the book offered Long a little extra money and the chance to interact with non-researchers interested in science. He is now considering another book idea.

Writing a book for a popular audience offers intellectual rewards and perhaps the opportunity for fame. But it also involves a lot of time, dedication and the willingness to re-examine one’s work with an eye towards popular consumption. Advance payments depend on market factors determined by publishers. There is no guarantee that the book will be profitable. And an author risks spending plenty of toil on work that might not make money or even be well received.

The process is taxing. “Nobody has ever finished a book and said to me in all my years in this business, ‘That was easier than I thought,’” says agent Alice Martell of the Martell Agency in New York City. To give the project its best chance at success, researchers who want to tell their stories to a broad audience need to find an appealing idea, and work with a team to craft the book.

SEEDS OF AN IDEA

Most popular books by scientists either link directly to their research or explore big topics in their field. An idea might come from noticing a gap in the market. As an avid reader, Daniel Davis recognized that there were many popular-science books on fundamental physics or evolutionary biology. But Davis, an immunologist at the University of Manchester, UK, had not seen one about his own research area. That void led to him to come up with the topic for his 2013 book, *The Compatibility Gene* (Penguin), which explores the diversity of ►

POPULAR SCIENCE

Get the word out

Writing science books for the public is rewarding, but can take a lot of time — and financial gain is uncertain.

► human immunity and its implications. To make it accessible to the public, he focused on the idea that the biggest genetic differences between humans are not in the genes that code for hair, eye or skin colour, but rather in the genes of our immune systems.

Scientists considering a book should seek advice from a literary agent early in the process, to avoid investing a lot of time in an idea that will not work. Because they know the business and what editors are looking for, agents can reshape an idea to be more attractive to a publisher, says Long. That is crucial, says Martell, because a prospective author needs to persuade a publisher that the idea has merit and that he or she is the best person to write it (see “The narrative arc of authorship”). It also helps to establish a following online — a social-media presence and a popular blog, for example, can show an agent or publisher that a researcher has a potentially promising readership.

SUBSTANCE WITH STYLE

Writers must cater to their audiences. Most readers of popular-science books are well educated, but are not in the field. If they are to spend US\$25 on a hardback book and invest the time to read 300 pages, they also want to

be entertained. “You really have to take the readers someplace that they couldn’t go otherwise, and it has to be a compelling read,” says Martell. Most non-fiction books, even those about science, rely on narrative storytelling — very different from the dry, expository style of journal articles or the chatty, informative writing suited to a blog or opinion piece.

The style should also be more literary than a researcher is probably used to, using a more conversational tone and perhaps including personal stories or jokes about the science, says Davis. Retelling how the science is done is as important as explaining the results, says Steven Pinker, a cognitive scientist at Harvard University in Cambridge, Massachusetts, whose best-selling books include *The Language Instinct* (Harper Perennial, 1994). To make stories come alive, Pinker spends a lot of time digging into the nitty-gritty of methods sections in papers.

It can be tricky to get away from jargon and still avoid talking down to intelligent readers. When Irene Pepperberg, a psychologist and research associate at Harvard, was working on her book *The Alex Studies* (Harvard University Press, 2002), which focused on her cognitive research on parrots, she tested her writing by



DAVID CARTER

Irene Pepperberg’s book about cognitive research in parrots caught the public’s imagination.

sending drafts of chapters to friends who had been to university but worked outside her field. “If they could follow it, then I was home free,” she says. Pepperberg says that parrot owners liked having a book that explained their pets’ behaviour, whereas scientists liked having a round-up of all the relevant literature.

Word choice and sentence rhythm are important, too. Writers need, for example, to vary sentence length and provide smooth transitions, says David Haskell, a biologist at Sewanee: the University of the South in Tennessee and author of *The Forest Unseen* (Viking Penguin, 2012). He compares writing to making music or cooking: the words need to sound and taste right. The richness of the writing also needs to transport readers to a place where they are not necessarily conscious of learning something, says Martell, who is Haskell’s agent. “It’s being able to pull other people in and making them want to read it.”

“It’s not about little literary flourishes. It’s about the squirrels or the bacteria,” Haskell says. “They need to be honoured with good words.”

SHAPE AND TIME

The book must have a structure with a clear beginning, middle and end. The big-picture narrative and pacing can be challenging for any author, especially one navigating a sea of research. That is where help from an editor is invaluable, says Haskell: for him, input about narrative structure “really took the book up several notches in quality”. Kevin Doughten, who was Haskell’s main editor and now works at Crown in New York City, says that the overall quality of Haskell’s writing was superb. So Doughten’s role was to find places where the pace of the book moved a little too quickly or where he felt that a reader’s brain might get tangled up. “Here are the parts where I need you to slow down a little bit or I need you to explain,” he advised.

INTO PRINT

The narrative arc of authorship

Writing a book involves more than putting pen to paper. Here are the six main steps in crafting a popular-science book.

- Find a big idea. You will need to convince an agent (and eventually a publisher) that your concept is a great idea for a book and that you are the best person to write it.
- Find an agent to serve as a liaison between you and the publisher, and to help you to navigate the publishing process. He or she can also help you to refine your idea, polish your book proposal and negotiate your publishing contract, and will be an advocate on your behalf. If you have a public profile, an agent might approach you; otherwise, you can talk to colleagues about their agents, or check the acknowledgements sections of books that you admire. Advances — upfront payments an author receives from the publisher that may be recouped with book sales — range widely. Writers might expect anything from US\$10,000 to \$80,000 for a niche topic, but a great idea with obvious appeal to the public could get much more. Your agent typically receives around 15% of total earnings, including both the advance and royalties. Other money may come from international rights and rights to different media; agents also get a portion of these earnings.
- Write a proposal. Editors want an idea

of what you are planning to write before you write it, so authors typically work with an agent to prepare a proposal with a compelling overview of the book, sometimes with one or more sample chapters and a chapter-by-chapter outline. It should include information about books on similar topics. Your agent will send your proposal to publishers and editors to gauge interest, which can take weeks or months.

● Write the book. After signing a contract, you can expect to have a year to write your book. During that period, you will probably want to send a few chapters to your editor early on to ensure that your expectations match.

● Revise the book. Once you hand it in, you might work with your editor to expand, prune and refine the book and ensure that the whole manuscript works. Accuracy is important: you will need to fact-check and index the book (or hire someone to help you). Then, over several months, copy editors and proofreaders will work through the accepted manuscript as you review changes.

● Market the book. Once it goes on sale, you will need to take part in publicity. You might do television and radio appearances, or give talks about your topic. Your publisher may expect you to blog or participate in social media, and you should let your contacts know about your book to help with sales. **S.W.**



Daniel Davis wrote about differences in genes of the human immune system, a topic that had received little attention in the popular-science market.

As a big project with multiple moving parts, writing a book can be all-consuming. Davis says that it is almost impossible to be a brilliant teacher, researcher and author all at the same time. While writing, he kept up his lab but did not teach. Most scientist-authors write books during a sabbatical or on leave while teaching is not in session. Pepperberg wrote *The Alex Studies* in three-hour blocks from 9 a.m. to noon each day, and spent the rest of her 13-hour workdays in her laboratory. The writing, she says, took the place of teaching and committee work during that time.

Once a book is published, authors have to get involved in marketing. They need to be able to talk about it to an average reader, and might be asked to do interviews on radio or television and give talks. Marketing can be gruelling. "There's just not enough hours in the day and not enough ATP in the body," says Haskell, who is currently taking a year of unpaid leave from teaching to promote his current book and to prepare his next one. For this year, his income comes from his book earnings and from fees for book-related speaking engagements.

Because of these demands, most scientist-authors advise other academic researchers not to start writing a book before they earn tenure. "I think it's a mistake until you're comfortable establishing a lab and getting grants," says Pinker. But Davis says that people with an enduring passion for book writing should just go ahead and dive in when offered the opportunity — or they might miss the chance.

MEASURING SUCCESS

Predicting which books will succeed and which will fail is never easy. Even if a book gets critical acclaim, there is no guarantee that the public will embrace it. Haskell recognizes that he is one of the lucky ones: his *The Forest Unseen* was a finalist for the 2013

Pulitzer Prize for General Non-Fiction, and won this year's US National Academy of Sciences book award. It has also sold well, says Haskell, both through mainstream booksellers and as a result of being adopted for use in academic courses.

But regardless of commercial success, the mental energy that goes into book writing can enrich a scientific career. Pinker sees writing as an extension of his academic research; it is like doing theoretical science with an audience. Davis has taken advantage of the time he spent thinking about 60 years of research and interviewing other scientists: since writing his book, he has used the ideas he came across to pull out new themes for his group's research. Seeing the big picture revealed new ways to focus on important questions in the field. For example, his laboratory is now looking at how different individuals' immune systems respond to various types of diseased cells.

Long's book has raised his public profile and led him to two unexpected opportunities. He is developing an 'Introduction to Robotics' course for the Great Courses, a company in Chantilly, Virginia, that sells DVD-based teaching materials. Long has also formed a research collaboration with Josh Bongard, a cognitive researcher at the University of Vermont in Burlington whom he met after Bongard reviewed his book for the magazine *New Scientist*.

Book publishing has always been a risky business. "The amount of work isn't proportional to the pay-off," says Laura Wood of Fine Print Literary Management in New York City. It is not about the money, Haskell emphasizes. "Irrepressible love of language and science," he says, "is the only good reason I can think of to set pen to paper." ■

Sarah Webb is a freelance writer based in Chattanooga, Tennessee.

RELOCATION

International impact

Scientists who move countries tend to publish in higher-impact journals than those who remain at home, a study finds (C. Franzoni *et al. Econ. Lett.* <http://doi.org/p68>; 2013). The authors asked about the relocation history of 14,299 researchers of all career stages in biology, chemistry, Earth and environmental sciences and materials science in 16 nations. Looking at papers published in 2009, the team found that scientists who were living in countries other than the ones they had been living in at age 18 published in journals with impact factors an average of 1.07 points higher than scientists who stayed put. Moving may help scientists to find work settings where they can maximize their potential, says co-author Chiara Franzoni, an economist at the Polytechnic University of Milan in Italy.

DIVERSITY

Inequalities at work

Women of colour comprised 5.7% of US science, technology, engineering and maths (STEM) academic faculty members with doctorates in 2010, a report says; white men made up 58%. *Accelerating Change for Women Faculty of Color in STEM* adds that the low numbers and restricted advancement of minority women on STEM faculties limit innovation and role models. It notes that university leadership should value diversity, but women of colour must cut time spent on committee service and mentoring, learn how job duties count towards tenure and pay rises, and welcome help, says Barbara Gault, co-author of the report and vice-president of the non-profit Institute for Women's Policy Research in Washington DC.

RECRUITMENT

Florida hiring push

The University of Florida in Gainesville is recruiting 100 researchers in 16 fields including neurology, global health, plant genomics, metabolomics and drug discovery. The move is part of a strategy to become a leading research university, says spokesman Chris Moran. The institution has received US\$15 million in state funds, which it will match with privately raised money to create 107 endowed seats that could be occupied by new recruits or existing faculty members. The money will also support the construction of research facilities and other initiatives.

PLANETARY DEFENCES

Credit where credit is due.

BY S. R. ALGERNON

“We’ve lost another mining platform,” said Overseer Kleeg. “A shaped antimatter charge breached the deflector field. Earther slaves don’t have that sort of technology. I want you to find out if one of the major powers is conspiring against us.”

Security Chief Vig’lah clicked his front pincers in frustration. “I would have detected any plot, Overseer. We should not overlook the possibility that the Earthers acquired the weapon themselves. You’ve seen their video transmissions. For a race incapable of interstellar travel, they are particularly sophisticated in their countermeasures against invasion.”

“But those transmissions were fiction.”

“Nevertheless, Overseer, we have taken precautions. We set up quarantine zones against microbes. We kept our weapons and written materials under lock and key — including all cookbooks. We sealed off the mother ship and confiscated all laptop computers. Earthers must have other tricks that we did not anticipate. You have my word, Overseer. I will uncover them.”

“Success at last,” said Vig’lah. He stridulated with joy.

“We have lost two more platforms,” said Kleeg. “Get to the point.”

“Some of our subcontractors use antimatter power for their tractor beams. I noticed that requisitions for replacement parts have increased lately at several construction sites. The Earthers could have used parts from these sites to build a bomb.”

“Earthers have broken into our construction sites?” Kleeg’s antennae retracted.

“No, no,” said Vig’lah. “The security perimeter was not at fault. We found a note to one of the construction managers in the debris of one of the platforms. It says: ‘Little Annie is only seven years old, and she will die of — this word is untranslated, but it must be some local malady — unless we can supply power to the abandoned hospital near crater 437. Please accept 2,000 galactic credits in exchange for one magnetic bottle of antimatter. If you help us save dear Annie’s life, we will be overjoyed. Production will surely increase by 15% next cycle.’”

“They are a conquered people,” said Kleeg. “How did they

get access to galactic currency? Who would have paid them?”

“You have my word, Overseer —”
“Enough promises. Get to work.”

“Overseer, I have news,” said Vig’lah. He spoke solemnly. Three platforms had fallen since his last audience with Kleeg, so he could not gloat. He carried a display tablet in his forward pincers.

“Very well. Speak.”

“After scanning the planet for the RF signature of galactic currency, we found 20,000 credits in an impact crater on a remote coastline. Officially, the impact was from an obsolete surveillance satellite, but I believe that the money had been hidden inside the satellite prior to its descent.”

“Only the Communications Guild has access to orbital satellites. Could they really have turned against us?”

“They would be a formidable enemy,” said Vig’lah. “However, I have made discreet enquiries. It turns out that a delegation from the Communications Guild visited the planet one cycle ago. Their visit coincided with an entertainment broadcast by the natives. We did not think much of it at the time.”

“Native broadcasts? Isn’t that dangerous?”

“It boosted morale. We had our censors vet every transmission. Earther entertainment can be quite outlandish, and we thought —”

“Is that the transcript you’ve got there?”

“Well, yes, but please understand —”

Kleeg snatched up the transcript. He read with a growing sense of foreboding.

Greetings. I am being called the Esteemed Denzel Gatez Starbuck, last in line to be Prince of Planet Earth. In times past, my family amassed large fortune in the Earth Province of West Prometheus, valued at 1.3 MILLION galactic credits. Unfortunately, during Invasion of my Planet, the monies were buried by an atomic blast. If you could provide us with 20,000 galactic credits for the equipment to recover the vault, we would gladly remit to you 50% of the riches within. Have a joyous cycle.

Kleeg could not bear to read any further.

“Tell me,” he said, “that you can put an end to these transmissions.”

“Alas, the natives have used their proceeds to buy access to the interstellar net from

the Communications Guild. Fortunately, their encryption is primitive, so we can tell what they are up to. Some of their messages depict members of the starfaring civilizations in scandalous states of undress. Some of them offer to sell land — on Jupiter, of all places. Inbound financial transactions have increased by three orders of magnitude.

These are heavily encrypted, but I suspect a substantial influx of credits to Earth. The loyalty of the independent contractors can no longer be assured.”

“You are Security Chief. Surely you can find some pretext for shutting down planetary communications.”

“I’m afraid not. The recent loss of mining platforms has substantially reduced the value

of your collateral. My sources say that the mining venture will soon be bankrupt, and that a consortium of natives is set to buy the remaining platforms.”

“I have heard nothing from High Command. How do you know this?”

“Apparently, the natives are in need of a security chief. They value my knowledge of the local mining operations, and they appreciate it that I allowed them to broadcast their movies back when they worked for you. I am here to offer my resignation.”

Kleeg relaxed his internal air bladders. *I should have known, he thought, that the Earthers would not make things easy.*

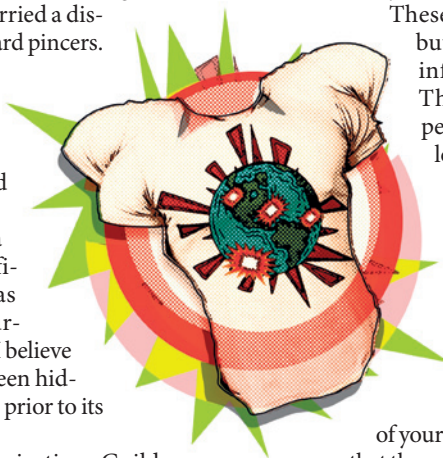
“What will you do?” asked Vig’lah. “The natives probably have no use for an overseer.”

Maybe not, thought Kleeg, but there’s still time to reassemble the invasion fleet. After Vig’lah had left, Kleeg tracked down one of the captured laptops. Somewhere in its memory, Kleeg knew, there was an Earther trick that would save the expedition from ruin.

Only 18 standard galactic cycles to go! Pledge 100 credits and win an “I HELPED CONQUER THE EARTH” T-Shirt!

The Earthers can keep Vig’lah — and their fake princes, too, thought Kleeg. I don’t need them, not when I’ve got QuickConquer. ■

S. R. Algernon studied fiction writing and biology, among other things, at the University of North Carolina at Chapel Hill. He currently lives in Singapore.



JACEY

➔ NATURE.COM

Follow Futures:

@NatureFutures

go.nature.com/mtoodm